DOCUMENT RESUME

ED 438 326	TM 030 643
AUTHOR	Dings, Jonathan
TITLE	KIRIS Accountability Cycle 2 Technical Manual. Based on the Analysis of Data from the 1992-93 through 1995-96 School Years.
INSTITUTION	Kentucky State Dept. of Education, Frankfort.
PUB DATE	1997-05-00
NOTE	337p.; For the Cycle 1 Technical Manual, see TM 030 642.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Descriptive (141)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Academic Achievement; *Accountability; *Educational Assessment; Educational Change; Elementary Secondary Education; *Psychometrics; Scoring; State Programs; *State Standards; Test Construction; Test Use; Testing Problems; *Testing Programs; Validity
IDENTIFIERS	Kentucky; *Kentucky Instructional Results Information System; Test Specifications

ABSTRACT

This technical manual provides information about the technical characteristics of the Kentucky Instructional Results Information System (KIRIS), the assessment component of the Kentucky Education Reform Act Accountability System. The time period covered is Accountability Cycle 2, which spanned the school years 1992-93 through 1995-96. Sections of the manual, which traces the development of the tests and discusses their psychometric characteristics, are: (1) "Context"; (2) "Test Development and Item Analysis"; (3) "Administration, Scoring, and Reporting"; (4) "Generalizability"; and (5) "Appendices." Thirteen appendixes contain supplemental information and some forms used in KIRIS. (Contains 109 tables and 11 figures.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 438 326

KIRIS ACCOUNTABILITY CYCLE 2 TECHNICAL MANUAL

BASED ON THE ANALYSIS OF DATA FROM THE 1992-93 THROUGH 1995-96 SCHOOL YEARS

[illegible]

A. Arrastia

2A [illegible]

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM030643

Produced by the Kentucky Department of Education
Wilmer S. Cody, Commissioner
May, 1997

BEST COPY AVAILABLE

DEDICATION

As is all of school reform in Kentucky, this manual is dedicated to our children. May their dreams never be fettered by a lack of learning.

ACKNOWLEDGEMENTS

This manual has benefited from the work and advice of many individuals. While we are grateful for this assistance, sole responsibility for the accuracy and completeness of this manual rests with the author. Table A presents an alphabetical list of those involved with this manual and their affiliations.

AUTHORSHIP. The editor and primary author of this technical manual is Jonathan Dings. Others whose contributions to individual chapters should be recognized are: William Insko, for writing sections of chapters 6, 9, 13, and 14, and conducting many of the analyses reported in this manual; Sue Rigney, Starr Lewis, Amy Awbrey, and Sarah Kennedy, for authorship of chapter 12; D. Clayton Smith, for writing a section of chapter 15 and conducting analyses pertinent to several chapters; and, Robert Wetter, for authorship of Chapters 10 and 11. The design of this technical manual and much of its text were based on the KIRIS Accountability Cycle I Technical Manual, for which the primary authors were Jonathan Dings, Brian Gong, and Neal Kingston. Portions of several chapters, most notably chapter 5, are excerpted with minor revision from longer reports published by the Kentucky Department of Education or Advanced Systems.

DATA ANALYSIS. Data analyses and tables for this manual were produced by Julie Ashworth, Diane Bailey, Charles DePascale, Jonathan Dings, Brian Gong, C. L. Huling, William Insko, Stuart Kahl, Mary Kino, D. Clayton Smith, C. Scott Trimble, and Robert Wetter.

ADVICE AND REVIEW. Throughout the second accountability cycle, Richard Hill regularly designed and carried out psychometric analyses critical to the implementation of the testing and accountability programs. Scott Trimble provided significant input on all aspects of the testing and accountability programs. On an on-going basis, we received psychometric advice from Robert Mislevy and the members of our National Technical Working Group – Edward Haertel, Skip Kifer, Paul LeMahieu, the late Brenda Loyd, Joseph Pedulla, John Poggio, and David Wiley.

Amy Awbrey, Jonathan Dings, Betty Edwards, Lisa Ehrlich, Brian Gong, William Insko, Ed Reidy, Sarah Kennedy, and Scott Trimble served as technical reviewers. Ann Bartosh, Nijel Clayton, Agnes Durbin, Mary Alice Harmon, Joanne Jensen, Tricia Kerr, Christie Maloney, Jim McMann, Jo Ann Mosier, Renee Murray, Arthur Patterson, Fran Salyers, Rhonda Sims, and Larry West reviewed the test development sections.

OTHER SUPPORT. Teri Eads, Otis McCreary and Tracy Tingle provided clerical support.

Table A
People Acknowledged and Their Affiliations

Julie Ashworth, Former Research Assistant/Data Analyst, Allegiant Management Corporation
Diane Bailey, Test Developer, ASME
Ann Bartosh, Mathematics Consultant, KDE
Amy Awbrey, Writing Portfolio Consultant, ASME
Nijel Clayton, Social Studies Consultant, KDE
Charles DePascale, Former Supervisor of Data Analysis and Scoring, ASME
Jonathan Dings, Group Leader, Psychometric Analysis, Allegiant Management Corporation
Agnes Durbin, Branch Manager, Curriculum and Assessment Development, KDE
Betty Edwards, Director of Curriculum and Assessment Development, KDE
Teri Eads, Secretary, KDE
Lisa Ehrlich, Kentucky Project Director, ASME
Brian Gong, Associate Commissioner for the Office of Curriculum, Assessment, and Accountability, KDE
Edward Haertel, Professor, Stanford University
Mary Alice Harmon, Writing Portfolio Consultant, KDE
Richard Hill, President, ASME
C. L. Huling, Performance Events Coordinator, ASME
William Insko, Database Analyst, KDE
Joanne Jensen, Research Associate, WestEd
Stuart Kahl, Vice President, ASME
Sarah Kennedy, Alternate Portfolio Director, Systems Change
Tricia Kerr, Science Consultant, KDE
Edward "Skip" Kifer, Professor, University of Kentucky
Neal Kingston, Vice President, ASME
Mary Kino, Director, Measurement and Data Analysis, ASME
Paul LeMahieu, Delaware Department of Education
Starr Lewis, Manager of Writing Portfolios, KDE
Brenda Loyd (deceased), Professor, University of Virginia
Christie Maloney, Director of Educational Programs, ASME
Otis McCreary, Database Operator, Allegiant Management Corporation
Jim McMann, Project Director, Data Recognition Corporation
Robert Mislevy, Distinguished Research Scientist, Educational Testing Service
Jo Ann Mosier, Manager of Mathematics Portfolios, KDE
Renee Murray, Language Arts Consultant, KDE
Arthur Patterson, Arts and Humanities Consultant, KDE
Joseph Pedulla, Professor, Boston College
John Poggio, Professor, University of Kansas
Ed Reidy, Deputy Commissioner, Learning Results Services, KDE
Sue Rigney, Director of Portfolio Initiatives, KDE
Fran Salyers, Middle Level Project Coordinator, KDE
Rhonda Sims, Humanities Manager, KDE
D. Clayton Smith, Research Assistant, Allegiant Management Corporation
Tracy Tingle, Secretary, KDE
C. Scott Trimble, Director of Assessment Implementation, KDE
Larry West, Science Consultant, KDE
Robert Wetter, Senior Systems Consultant, KDE
David Wiley, Professor, Northwestern University

CONTENTS

Dedication ... i
Acknowledgements ... ii
Contents ... iv

SECTION I: CONTEXT

1. Introduction ... 1-1
2. Background ... 2-1

SECTION II: TEST DEVELOPMENT AND ITEM ANALYSIS

3. Test Specifications ... 3-1
4. Test Development Process ... 4-1
5. Setting and Validating Performance Standards ... 5-1
6. Item Analysis ... 6-1

SECTION III: ADMINISTRATION, SCORING, AND REPORTING

7. Test Administration ... 7-1
8. Scoring ... 8-1
9. Equating and Linking ... 9-1
10. Reporting to Schools and Districts ... 10-1
11. Statewide Summary of Results ... 11-1
12. Portfolio Assessment: Scoring and Student Performance ... 12-1

SECTION IV: GENERALIZABILITY

13. Reliability ... 13-1
14. Validity - Related Evidence ... 14-1
15. Consequential Validity of KIRIS and the Kentucky Assessment Program ... 15-1

SECTION V: APPENDICES

SECTION I
CONTEXT

Evaluation of any program must be in terms of its goals and purposes. To this end, the first section of this technical manual describes the purpose and contents of this manual (Chapter 1) and the background of the testing program and accountability system (Chapter 2).

CHAPTER 1
INTRODUCTION

The purpose of this technical manual is to provide information about the technical characteristics of the Kentucky Instructional Results Information System (KIRIS), the assessment component of the Kentucky Education Reform Act (KERA) Accountability System. The time period covered is Accountability Cycle 2, which spanned the school years 1992-93 through 1995-96. While some parts of this manual may be used by educated laypersons, its intended audience is experts in psychometrics and educational research. It assumes working knowledge of measurement concepts such as reliability and validity, and statistical concepts of correlation and central tendency. For some chapters, the reader is presumed to have basic familiarity with advanced topics in measurement and statistics such as item response theory and factor analysis.

COMPREHENSIVENESS

Despite the many pages of tables and prose in this manual, it is by no means intended to provide all available detail about the testing program or accountability system. To do so would have made the technical manual prohibitively expensive. Instead, summaries of important studies are provided with references in addition to information unique to this manual. Some data about the first two years of Accountability Cycle 2 (1992-93 and 1993-94) are not presented in this volume, as they are available in KIRIS Accountability Cycle I Technical Manual.

NOT JUST A TEST

In reading this technical manual, it is critical to remember that the testing program does not exist in a vacuum. It is one part of a complex network intended to help schools focus their energies on dramatic improvement in student learning. That is, KIRIS is part of an integrated program of testing, accountability, and curricular and instructional support, coupled with wide-sweeping changes in school finance, governance, and organization. It can only be evaluated properly in this full context.

CHAPTER 2
BACKGROUND

To understand the Kentucky testing and accountability programs, one must understand the context from which they arose. This chapter describes the court case and ensuing executive and legislative actions that led to the current system.

KENTUCKY CONSTITUTION

The Kentucky Constitution, adopted in 1891, states that the "*General Assembly shall, by appropriate legislation, provide for an efficient system of common schools throughout the state.*"

COUNCIL FOR BETTER EDUCATION V. WILKINSON

In November 1985, the Council for Better Education, a nonprofit corporation formed by 66 school districts, seven boards of education, and 22 public school children sued the state for not providing an efficient system of education. In October 1988, Franklin County Circuit Court judge Roy Corns found for the plaintiffs. Seven capacities establishing a new legal framework for the school curriculum were set forth by the court as necessary for an adequate education:

> *Schools must provide (i) significant oral and written communication skills to enable students to function in a complex and rapidly changing civilization; (ii) sufficient knowledge of economic, social, and political systems to enable the student to make informed choices; (iii) sufficient understanding of government processes to enable the student to understand the issues that affect his or her community, state, and nation; (iv) sufficient self-knowledge and knowledge of his or her mental and physical wellness; (v) sufficient grounding in the arts to enable each student to appreciate his or her cultural and historical heritage; (vi) sufficient training or preparation for advanced training in either academic or vocational fields so as to enable each child to choose and pursue life work intelligently; and (vii) sufficient levels of academic or vocational skills to enable public school students to compete favorably with their counterparts in surrounding states, in academics or in the job market. Council for Better Education v. Wilkinson, NO. 85-CI-1759, slip op. at 4 (Franklin Cir. Ct., Oct. 14, 1988).*

In June 1989, the Kentucky Supreme Court directed the General Assembly to recreate and reestablish a new "efficient system of common schools" that complied with the Kentucky Constitution. The Court defined an efficient system of common schools as "an organization that provides a free and adequate education to all students throughout the state regardless of geographical location or local fiscal resources."

COUNCIL ON SCHOOL PERFORMANCE STANDARDS

In February, 1989, on his own initiative, Governor Wallace Wilkinson issued an executive order creating a twelve-member Council on School Performance Standards. Charged with determining what all students should know and be able to do and how learning should be assessed, the Council began its work by traveling across the Commonwealth conducting focus group interviews with business leaders, employers of graduates, parents, and educators, and asking them what a high school graduate in the year 2000 should know and be able to do. Information from the group interviews was used to construct a twenty-two minute telephone survey. More than 830 Kentucky residents of voting age were called and asked their opinions about what should be expected of future high school graduates.

The responses from the nearly 300 hours of telephone interviews were clear and consistent. While there was plenty of support for learning basic communications and mathematics skills, the most important result of learning identified was that students should be able to use their skills and knowledge in mathematics, science, social studies, and other disciplines in situations similar to what they will experience in the real world. In addition, Kentucky residents believed students should develop personal attributes such as self-sufficiency, self-confidence, and adaptability. They should learn to become responsible members of a family, work group, and community as part of their school learning experience. Also, Kentuckians believed students should learn to think and solve problems, make informed choices, and support their decisions. The information from the statewide survey was presented to five task forces of educators from across the state who were asked to advise the Council on possible learning goals for all students and ways in which these goals might be assessed.

In September, 1989, the Council on School Performance Standards produced the report *Preparing Kentucky Youth for the Next Century: What Students Should Know and Be Able To Do and How Learning Should Be Assessed* and presented it to the Curriculum Committee of the Legislative Task Force charged with creating Kentucky's new "efficient system of common schools." Six broad learning goals for all students were recommended with particular emphasis on what they should be able to do. These learning goals are presented in Table 2-1.

In addition, the Council recommended that the state launch a major effort to assess student performance beyond what can be measured by paper-and-pencil tests. It also was recommended that the state initiate long-range development efforts that support school reform in implementing the new learning goals.

KENTUCKY EDUCATION REFORM ACT

In 1990, the Council's recommendations were incorporated into House Bill 940, the Kentucky Education Reform Act (KERA), as a first step in redefining the school curriculum and providing what the courts required as an "adequate education for all students." The bill was signed by Governor Wallace Wilkinson on April 11, 1990 and became law on July 13, 1990.

With KERA, the General Assembly established the framework for a major revision of Kentucky's educational system. KERA required the establishment of learning goals for the educational system, provided a procedure by which those goals would be defined and assessed, and created a series of rewards and assistance to be associated with performance of schools on those assessments. Table 2-2 presents the text of one portion of the Kentucky Education

Reform Act, KRS 158.645 – *Capacities Required of Students in Public Education*, which is a modification of language used in the court decision.

TABLE 2-1
KENTUCKY'S SIX LEARNER GOALS

1. Students are able to use basic communication and mathematics skills for purposes and situations they will encounter throughout their lives.
2. Students shall develop their abilities to apply core concepts and principles from mathematics, the sciences, the arts, the humanities, social studies, practical living studies, and vocational studies to what they will encounter throughout their lives.
3. Students shall develop their abilities to become self-sufficient individuals.
4. Students shall develop their abilities to become responsible members of a family, work group, or community, including demonstrating effectiveness in community service.
5. Students shall develop their abilities to think and solve problems in a variety of situations they will encounter in life.
6. Students shall develop their abilities to connect and integrate experiences and new knowledge from all subject matter fields with what they have previously learned and build on past learning experiences to acquire new information through media sources.

TABLE 2-2
KRS 158.645 – CAPACITIES REQUIRED OF STUDENTS IN PUBLIC EDUCATION.

The General Assembly recognizes that public education involves shared responsibilities. State government, local communities, parents, students, and school employees must work together to create an efficient school system. Parents and students must assist schools with efforts to assure student attendance, preparation for school, and involvement in learning. The cooperation of all involved is necessary to ensure that desired outcomes are achieved. It is the intent of the General Assembly to create a system of public education which shall allow and assist all students to acquire the following capacities:

(1) Communication skills necessary to function in a complex and changing civilization;
(2) Knowledge to make economic, social, and political choices;
(3) Understanding of governmental processes as they affect the community, the state, and the nation;
(4) Sufficient self-knowledge and knowledge of his mental and physical wellness;
(5) Sufficient grounding in the arts to enable each student to appreciate his or her cultural and historical heritage;
(6) Sufficient preparation to choose and pursue his life's work intelligently; and
(7) Skills to enable him to compete favorably with students in other states.

Kentucky's new reform act, however, went beyond adopting six learning goals. Elements of the law that have a direct effect on the curriculum and what students learn include the following:

1. *Schools shall require students to demonstrate a high level of achievement on each of the six learning goals.*
2. *Schools shall increase their attendance rates, reduce drop-out and retention rates, reduce physical and mental barriers to learning, and increase students' successful transition to post-high school experiences.*
3. *The Council on School Performance Standards shall frame the six school goals in measurable terms that define expected student learning outcomes. Statewide committees of practitioners shall be involved in this task.*
4. *The State Board for Elementary and Secondary Education shall create and implement a statewide, primarily performance-based assessment program.*
5. *All schools in the Commonwealth shall be held accountable for the proportion of students who achieve the school goals at the expected level set by the State Board.*

These requirements of the Reform Act placed special conditions on how the Council was to frame the school goals in measurable terms and how it would redefine what students learn in school. First, since there would be high stakes accountability (known as rewards and sanctions for schools) and teachers would be "teaching toward the test," the test must be worth "teaching to." Therefore, the focus of learning must be on what students are able to do with their knowledge and skills. Second, since teachers are to be teaching toward goals which are framed in measurable terms, the goals must contain what is "essential" and most important for students to learn. Third, since the school is accountable for all students, the goals should describe what is expected for all students. It is important to note that the Council decided the academic expectations, which would define the goals in measurable terms, should not be defined as minimum competencies, but rather as standards of performance for educated citizens.

The Kentucky Educational Reform Act also called for the establishment of a committee of national testing experts to design the accountability system.

ACADEMIC EXPECTATIONS

To further define what was expected of students, in December 1991 the State Board adopted 75 statements, at that time called Valued Outcomes, describing what was expected of students. During the first three years of testing, more than 40 Valued Outcomes were incorporated into the new assessment system. In 1994 it became clear that the language used in the 75 Valued Outcomes did not clearly convey their intent to many important audiences, including parents, legislators, media commentators, and some educators. In addition, there were concerns both within and outside the Department of Education regarding the near-term feasibility of measuring Learner Goal 3 (self-sufficiency) and Learner Goal 4 (responsible group membership). To address both sets of concerns the Kentucky Department of Education (KDE) met with or received comments from more than 175 people including parents, educators, students, representatives of business and government, clergy, and non-affiliated citizens. Comments were received from both critics and advocates of Kentucky's education reform efforts.

The result of this process was the establishment of 57 academic expectations that describe what Kentucky students should be able to know and do when they graduate from high school. Table 2-3 presents the 12 academic expectations (the original outcomes 5-9 were combined into one academic expectation) related to Goal 1 – Communication Skills. Table 2-4 presents the 37 expectations (the original outcomes 16 and 21 were combined) related to Goal 2 – Core Academic Concepts. Table 2-5 presents the five expectations for Goal 5 –Think and Solve Problems, and Table 2-6 presents the three expectations for Goal 6 – Integrating Knowledge.

TABLE 2-3
ACADEMIC EXPECTATIONS FOR GOAL 1: COMMUNICATION SKILLS

1. Students use reference tools such as dictionaries, almanacs, encyclopedias, and computer reference programs, and reference tools such as interviews and surveys to find the information they need to meet specific demands, explore interests, or solve specific problems.
2. Students make sense of the variety of materials they read.
3. Students make sense of the variety of things they observe.
4. Students make sense of the various messages to which they listen.

5-9. Students use mathematical ideas and procedures to communicate, reason, and solve problems.

10. Students organize information through development and use of classification rules and systems.
11. Students write using appropriate forms, conventions, and styles to communicate ideas and information to different audiences for different purposes.
12. Students speak using appropriate forms, conventions, and styles to communicate ideas and information to different audiences for different purposes.
13. Students make sense of ideas and communicate ideas with the visual arts.
14. Students make sense of ideas and communicate ideas with music.
15. Students make sense of ideas and communicate ideas with movement.
16. Students use computers and other kinds of technology to collect, organize, and communicate information and ideas.

TABLE 2-4
ACADEMIC EXPECTATIONS FOR GOAL 2: CORE ACADEMIC CONCEPTS

SCIENCE

1. Students understand scientific ways of thinking and working and use those methods to solve real-life problems.
2. Students identify, analyze, and use patterns such as cycles and trends to understand past and present events and predict possible future events.
3. Students identify and analyze systems and the ways their components work together or affect each other.
4. Students use the concept of scale and scientific models to explain the organization and functioning of living and nonliving things and predict other characteristics that might be observed.
5. Students understand that under certain conditions nature tends to remain the same or move toward a balance.
6. Students understand how living and nonliving things change over time and the factors that influence the changes.

MATHEMATICS

7. Students understand number concepts and use numbers appropriately and accurately.
8. Students understand various mathematical procedures and use them appropriately and accurately.
9. Students understand space and dimensionality concepts and use them appropriately and accurately.
10. Students understand measurement concepts and use measurements appropriately and accurately.
11. Students understand mathematical change concepts and use them appropriately and accurately.
12. Students understand mathematical structure concepts including the properties and logic of various mathematical systems.
13. Students understand and appropriately use statistics and probability.

SOCIAL STUDIES

14. Students understand the democratic principles of justice, equality, responsibility, and freedom and apply them to real-life situations.
15. Students can accurately describe various forms of government and analyze issues that relate to the rights and responsibilities of citizens in a democracy.
16. Students observe, analyze, and interpret human behaviors, social groupings, and institutions to better understand people and the relationships among individuals and among groups.
17. Students interact effectively and work cooperatively with the many ethnic and cultural groups of our nation and world.
18. Students understand economic principles and are able to make economic decisions that have consequences in daily living.
19. Students recognize and understand the relationship between people and geography and apply their knowledge in real-life situations.
20. Students understand, analyze, and interpret historical events, conditions, trends, and issues to develop historical perspective.
21. (Incorporated into 2.16)

ARTS AND HUMANITIES

22. Students create works of art and make presentations to convey a point of view.
23. Students analyze their own and others' artistic products and performances using accepted standards.
24. Students have knowledge of major works of art, music, and literature and appreciate creativity and the contributions of the arts and humanities.
25. In the products they make and the performances they present, students show that they understand how time, place, and society influence the arts and humanities such as languages, literature, and history.
26. Though the arts and humanities, students recognize that although people are different, they share some common experiences and attitudes.
27. Students recognize and understand the similarities and differences among languages.
28. Students understand and communicate in a second language.

PRACTICAL LIVING

29. Students demonstrate skills that promote individual well-being and healthy family relationships.
30. Students evaluate consumer products and services and make effective consumer decisions.
31. Students demonstrate the knowledge and skills they need to remain physically healthy and to accept responsibility for their own physical well-being.
32. Students demonstrate strategies for becoming and remaining mentally and emotionally healthy.
33. Students demonstrate the skills to evaluate and use services and resources available in their community.
34. Students perform physical movement skills effectively in a variety of settings.
35. Students demonstrate knowledge and skills that promote physical activity and involvement in physical activity throughout their lives.

VOCATIONAL STUDIES

36. Students use strategies for choosing and preparing for a career.
37. Students demonstrate skills and work habits that lead to success in future schooling and work.
38. Students demonstrate skills such as interviewing, writing resumes, and completing applications that are needed to be accepted into college or other postsecondary training or to get a job.

TABLE 2-5
ACADEMIC EXPECTATIONS FOR GOAL 5: THINK AND SOLVE PROBLEMS

1. Students use critical thinking skills such as analyzing, prioritizing, categorizing, evaluating, and comparing to solve a variety of problems in real-life situations.
2. Students use creative thinking skills to develop or invent novel, constructive ideas or products.
3. Students organize information to develop or change their understanding of a concept.
4. Students use a decision-making process to make informed decisions among options.
5. Students use problem-solving processes to develop solutions to relatively complex problems.

TABLE 2-6
ACADEMIC EXPECTATIONS FOR GOAL 6: INTEGRATE KNOWLEDGE

1. Students connect knowledge and experiences from different subject areas.
2. Students use what they already know to acquire new knowledge, develop new skills, or interpret new experiences.
3. Students expand their understanding of existing knowledge by making connections with new knowledge, skills, and experiences.

SELECTING A CONTRACTOR

In 1991 the Kentucky Department of Education awarded a contract to Advanced Systems in Measurement and Evaluation, Inc., to assist KDE in designing and implementing the assessment and accountability system. The contract period was 1991-96. Toward the conclusion of the first contract period, the Kentucky Department of Education and the Kentucky Board of Education issued a Request for Proposals (RFP) October 9, 1995, for a contractor to assist the Department of Education implement the assessment system for the period 1996-2000.

The KIRIS RFP National Advisory Panel, consisting of seven persons representing expertise in measurement, assessment, state-wide testing, and instruction were recruited to assist in formulating the RFP. The National Advisory Panel also had the responsibility of evaluating the proposals received in response to the RFP.

The members of the KIRIS RFP National Advisory Panel, with their affiliations, are given below.

Raymond Burse, Chair
Senior Counsel for Transactions
G.E. Appliances, Louisville, KY
Past President, Kentucky State University

Ed Roeber, Co-Chair
Director, Student Assessment Program,
Council of Chief State School Officers,
Past Director, Michigan Educational Assessment Program

Edward "Skip" Kifer
Professor, University of Kentucky

John Poggio
Professor, University of Kansas
Director, Development of Kansas
Assessment Programs

Warren Simmons
Executive Director, Philadelphia Education
Fund, Philadelphia, PA

Brenda Loyd (Deceased December 1995)
Professor, University of Virginia
President, National Council on
Measurement in Education

Andy Porter
Professor, and Director, Wisconsin Center for
Education Research,
University of Wisconsin - Madison,
Chair, U.S. Department of Education,
Advisory Council on Education Statistics

A panel of 13 Kentucky citizens representing teachers, school administrators, and parents was also formed to evaluate the proposals. The members of the KIRIS Proposal Evaluation Panel-Kentucky, with their affiliations, are given below.

Terry Williams
Teacher, Panel Co-Chair
Tichenor Middle School
Erlanger-Elsmere Ind. Schools

Ken Draut
District Assessment Coordinator, Panel Co-Chair
Henry County Schools

Bill Caudill
Superintendent
Jenkins Independent Schools

Lue Cole
Teacher
South Laurel High School,
Laurel County Schools

Susan Cook
General Director of Federal
Programs & Testing and
District Assessment Coordinator,
Covington Independent Schools
and member, Kenton County School Board

Jessie Daniels
Coordinator, Resource Development
Jefferson County Public Schools

Gerald Hutchins
Associate Director of Research
Fayette County Schools

JoAnn Johnson
Pharmacist
Caldwell County

Karen Jones
Vice President, Kentucky PTA
Pulaski County

Rick McHargue
Principal/School Council Chair
Crestwood Elementary School,
Oldham County Schools

Doug Mings
District Assessment Coordinator
Hopkins County Schools

Ben Oldham
Professor
Georgetown College

Roland Haun
Executive Secretary
Kentucky Association of School Superintendents
(Fort Knox)

One proposal was received that addressed all components of the RFP (Accountability assessment, Non-Cognitive assessment, National Reference assessment, and Continuous assessment). This proposal was submitted by Advanced Systems, with subcontractors Data Recognition Corporation (DRC), Far West Laboratories (FWL), Human Resources Research Organization (HumRRO), Educational Testing Service (ETS), CTB/McGraw-Hill, Kentucky Systems Change Project (UK), National Computer Systems (NCS), Bayless & Associates.

Four other proposals were received for the National Reference assessment component of the RFP. Bidders were, CTB/McGraw-Hill (independently), Harcourt Brace Education Measurement, MESA Associates, and Riverside Publishing Company.

The ratings of the National Panel are given below.

Bidder (Advanced Systems and subcontractors)	Technical Soundness (350 points maximum)	Management Operation Plan (225)	Personnel, Staffing, and Organization (225)	Coherence Across Four Assessment Areas (50)	Cost (200)	TOTAL (1050)
Accountability Assessment	325	210	195	50	200	980
Noncognitive Assessment	320	210	215	50	200	995
National Reference Assessment	300	205	195	50	170	920
Continuous Assessment	275	205	180	50	200	910

*Note: The National Panel evaluated each component separately.

National Reference Assessment Bidder	Technical Soundness (350 points maximum)	Management Operation Plan (225)	Personnel, Staffing, and Organization (225)	Cost (200)	TOTAL (1000)
CTB/McGraw -Hill	300	205	195	170	870
Harcourt Brace Educational Measurement	325	205	190	200	920
MESA Associates	50	0	25	*	75
Riverside Publishing	275	160	170	100	705

*Note: Not rated, in accordance with RFP specification that a proposal must receive at least 70% of possible points in evaluation of first three areas of technical proposals in order for cost proposal to be considered .

The ratings of the State Panel are given below.

Total Package Bidder	Technical Soundness (350 points maximum)	Management Operation Plan (225)	Personnel, Staffing, and Organization (225)	Coherence Across Four Assessment Areas (50)	Cost (200)	TOTAL (1050)
Advanced Systems and partners	300	200	225	50	150	925

*Note: The state panel evaluated the entire proposal as a total package.

National Reference Assessment Bidder	Technical Soundness (350 points maximum)	Management Operation Plan (225)	Personnel, Staffing, and Organization (225)	Cost (200)	TOTAL (1000)
CTB/McGraw -Hill	300	200	220	180	900
Harcourt Brace Educational Measurement	300	190	190	200	880
MESA Associates	0	0	0	NA*	NA*
Riverside Publishing	290	190	200	100	780

*Not rated, in accordance with RFP specification that a proposal must receive at least 70% of possible points in evaluation of first three areas of technical proposals in order for cost proposal to be considered.

Based on these evaluation and the recommendations from the two independent panels, the contract for the 1996-2000 period was awarded to Advanced Systems and its subcontractors.

THE ACCOUNTABILITY SYSTEM

KIRIS data are used to hold schools accountable. This is done to heighten public attention in order to focus schools on helping their students achieve the high standards set by committees of Kentucky educators. Inherent in the accountability system is the recognition that the massive changes in instruction and learning will require many years to be achieved. Although there is no definite schedule, many have spoken of this as an effort that will take at least 20 years.

In order to keep schools focused throughout this lengthy period, a system of setting subgoals was developed. To reflect the distinctive starting points of different schools and their different populations, a unique improvement goal is calculated for each school to keep it moving in the right direction. Data from one testing period form the operational basis of implementing the policy decisions about setting each school's improvement goal.

KERA called for the integration of all information collected into three statistics for each school in the state. First, a ***baseline***, or initial accountability index, is computed. From this baseline, an ***improvement goal*** (in previous publications the improvement goal was referred to as the "threshold") is set. Each school's improvement goal represents a 10% improvement against the difference between its baseline and 100. Finally, a school's ***accountability index*** is calculated and compared to its improvement goal. For example, if a school's baseline is 30, the difference between its baseline and 100 is 70. Ten percent of 70 is 7. Thus, the school's improvement goal is 30+7, or 37.

Specifically, for the first accountability cycle, 1991-92 testing data provided the baseline against which the combined results of 1992-93 and 1993-94 testing were compared. Similarly, for second accountability cycle, the 1992-93 and 1993-94 testing data provided a new baseline against which 1994-95 and 1995-96 results were compared.

MEASURES. In addition to the assessment data collected through **on-demand testing, performance events,** and **portfolios** (for descriptions of each testing method, see Chapter 3, Test Specifications), KERA called for the accountability system to include certain noncognitive indicators of school success:

- attendance (used in all grades)
- retention (used in grades 4-12)
- drop-out rate (used in grades 7 through 12 only)
- transition to adult life (used in grade 12 only)

As with the cognitive areas, an advisory committee was created to ensure the input of Kentucky educators in the design of this component of the accountability system. Using the criteria called for by KERA as its foundation, the committee guided the development of the definitions of each variable and made suggestions for alterations in the system for future years.

EXTERNAL TECHNICAL ADVISORY PANEL

Two major studies of KIRIS were released in 1995. One study was sponsored by the Kentucky Institute for Education Research, the other by the Office of Education Accountability. Criticisms of KIRIS in both studies led to the creation of an external panel of nationally recognized measurement experts to advise the Kentucky Department of Education in implementing substantial changes in several aspects of the testing program. The names and affiliations are given in Table 2-10 of persons who were members of the National Technical Working Group during Accountability Cycle 2. Improvements to the testing program are reflected in this technical manual, and noted in Table 2-11 with reference to recommendations made in the study sponsored by the Office of Education Accountability.

TABLE 2-10
NATIONAL TECHNICAL WORKING GROUP

Name	Affiliation
Edward Haertel	Stanford University
Edward Kifer	University of Kentucky
Paul LeMahieu	University of Delaware, Delaware Dept. of Education
Brenda Loyd (deceased)	University of Virginia
Joseph Pedulla	Boston College
John Poggio	University of Kansas
David Wiley	Northwestern University, New Standards Project

TABLE 2-11
IMPROVEMENTS TO KIRIS IMPLEMENTED AFTER RECOMMENDATIONS OF OEA-SPONSORED PANEL

1995 Recommendation	Solution Implemented By Spring, 1997 KIRIS Test Administration
Develop and implement an ethical code of conduct	Code of Ethics was distributed to all educators in *Kentucky Teacher* (August, 1995), and incorporated as a regulation (703 KAR 4:110) in 1996.
Expand validation work	Since 1991-95, validation work has been expanded considerably by KDE, its contractor/subcontractor, and other researchers, with a strong emphasis on construct and consequential validity.
Strengthen equating design	Improved equating procedures were implemented beginning in Accountability Cycle 2 (1992-93 to 1995-96).
Re-establish performance standards	Existing standards for reading, mathematics, science and social studies were validated in October and November, 1995. Standards were re-established for arts & humanities and practical living/vocational studies.
Field test all score report forms	KDE field tested reports with five different groups of report users: (1) Superintendents, (2) District Assessment Coordinators, (3) Principals and Guidance Counselors, (4) Teachers, and (5) Parents.
Provide alternative explanations of results	Facilitated by contractor, the National Technical Working Group and Policy Advisory Committee review pertinent score release documents, including consideration of alternative explanations of results.
Establish routine auditing procedures for technical procedures, large score gains	National Technical Working Group reviews technical procedures. Auditing procedures for allegations of violations of Code of Ethics include the potential for a formal investigation. Schools with unusual patterns of score gains can have portfolio scores audited.
Include multiple-choice items	Multiple-choice items were pretested in 1995-96, and are included with open-response items in this year's testing. Multiple-choice will be included as part of baselines for Accountability Cycle 4 (1996-2000).
Improve documentation	The KIRIS Accountability Cycle I Technical Manual (about 300 pages in length) was issued in July, 1995; a similar report for Cycle 2 has been issued. Several other technical reports have been issued, including standards validation, standard setting, and portfolio score audit results, as well as Core Content.
Reconsider shift from content to process	KDE has examined open-response test content and process requirements. Content Guidelines were disseminated in fall 1994. Revised Core Content documents were presented in public meetings throughout the state beginning in fall 1995; final versions were distributed in July, 1996. Open-response and multiple-choice test development guidelines underscore the importance of content as well as application, with multiple-choice being targeted to enhance content coverage.
Portfolios should not be part of accountability system	Portfolios were designed to have a positive impact on instruction, require students to produce several pieces of work as evidence of meeting Kentucky standards, and directly reflect classroom curriculum and instruction. The portfolio system includes professional development for teachers, and is a valued component of KIRIS.
Teachers should not score their own school's portfolios	Scoring audit procedures provide regular statewide estimates of local scoring accuracy and validate or correct large score gains for individual schools. Local scoring accuracy improved dramatically from 1993 to 1996. KDE will continue to provide schools with enhanced training materials to help improve scoring accuracy further. KDE strongly encourages schools to document scoring accuracy.

SECTION II

TEST DEVELOPMENT AND ITEM ANALYSIS

The quality of the inferences made from test scores depends on the quality of the test development and standard setting processes. Item analysis provides us with objective evidence regarding the success of those efforts. Chapter 3 presents the frameworks used to ensure that each year's set of test forms is parallel to those of other years. Chapter 4 describes the test development process used to ensure that all tests and items are of very high quality. Chapter 5 describes how standards were set or validated to describe the four Kentucky performance categories. Chapter 6 summarizes the results of item analyses.

CHAPTER 3
TEST SPECIFICATIONS

There are two primary reasons why test specifications are necessary. First, they help ensure that the test is measuring what it is intended to measure. Second, test specifications help ensure that across the four years of an accountability cycle, the samples of tasks upon which school success is judged are equivalent.

Two sets of specifications have been used to develop KIRIS assessments. A subset of the 57 academic expectations that explicate four of Kentucky's six learner goals forms one set of specifications. Because the law requires KIRIS to be like the National Assessment of Educational Progress (NAEP)[1], a modification of the NAEP content frameworks forms the second set of specifications. Based on these two frameworks, Kentucky's assessment program targets large concepts and themes; the connecting of ideas; and the application of knowledge, skills, and reasoning to solve problems.

In addition to the aforementioned specifications, Kentucky's curriculum framework provides benchmarks and further information about content, concepts, and context for questions on the KIRIS tests. To provide an additional level of detail, KDE disseminated draft detailed content guidelines for all KIRIS tests in fall 1994. While these guidelines were first used to build the 1994–95 assessments, retroactively the most recent tests in the first accountability cycle were checked against them for compatibility.

LEARNER GOALS

The genesis of Kentucky's six learner goals is presented in Chapter 2. Table 3–1 lists the four goals that are the basis of the assessment program. (In Table 2–1 these are referred to as Goals 1, 2, 5 and 6; this numbering scheme is followed throughout the manual to avoid confusion.) For the second accountability cycle, the tests concentrated on these four learner goals; Goals 3 and 4 are, by statute, not assessed by KIRIS.

TABLE 3–1
KENTUCKY'S FOUR LEARNER GOALS THAT ARE MEASURED IN KIRIS

1. Students are able to use basic communication and mathematics skills for purposes and situations they will encounter throughout their lives.
2. Students shall develop their abilities to apply core concepts and principles from mathematics, the sciences, the arts, the humanities, social studies, practical living studies, and vocational studies to what they will encounter throughout their lives.
5. Students shall develop their abilities to think and solve problems in a variety of situations they will encounter in life.
6. Students shall develop their abilities to connect and integrate experiences and new knowledge from all subject matter fields with what they have previously learned, and build on past learning experiences to acquire new information through media sources.

---

[1] Although the law requires the transitional KIRIS to be NAEP-like, it does not define the characteristics of NAEP that should be embodied in KIRIS. Since the law also requires KIRIS to be primarily performance-based, KDE has defined NAEP-like to mean having substantially similar content frameworks.

ACADEMIC EXPECTATIONS

The history of the 57 academic expectations is described in Chapter 2. Demonstrating that these expectations are met requires students to show that they can use skills and knowledge in situations they might experience in life (e.g., writing a business letter, producing a report, solving a problem, or choosing and supporting the best course of action). That is, an academic expectation is a statement of student ability to demonstrate consistent quality performances on authentic tasks related to a skill area, core concept, or thinking process. Consistent quality performance means a high level of performance on a variety of tasks over time.

At this time, the statewide assessment does not reflect all fifty-seven academic expectations. Because the scope of reform was so comprehensive, it would have been unfair to hold all schools in Kentucky immediately responsible for the academic expectations which require special programs or facilities, e.g., foreign language. Additionally, the Kentucky General Assembly determined that the academic expectations in Learner Goals 3 and 4 (self-sufficiency and responsible group membership) should not be assessed at this time. Therefore, the KIRIS assessment does not reflect any academic expectations from those goals. Likewise, the academic expectations found in goals 5 and 6 have not been explicitly targeted for assessment (though many of the performance and portfolio items require the skills identified under these goals).

Tables 3–2 through 3–5 show which academic expectations were targeted with the open-response tests in each grade for 1994-95 and 1995-96; academic expectations for prior years are presented in the KIRIS Accountability Cycle I Technical Manual. Tables 3–6 through 3–9 present similar information based on the performance events tests. Tables 3–10 through 3–12 note the distribution of open-response items across academic expectations. During the 1995–96 testing year, multiple-choice and additional open-response items were included in a pre-test section in grades 4, 8, and 11. These items were not included in the accountability index and were not reported to schools. The intent of this section was for large scale item field testing.

Because the open-response questions tend to be broad and often belong to multiple categories (i.e., address multiple academic expectations), totaling the numbers of items in the categories within a major subject area will usually yield a larger number than the actual count of items. This "double counting" of questions might lead to vastly disproportionate numbers of items in an area. This is especially true in science, where large numbers of items fall into the patterns and systems categories. (Some would agree that science is the study of patterns.) Data and procedures (operations) in mathematics also tend to "capture" many items. Some categories are considerably overlapping from one another by definition, e.g., measurement and data, democratic principals and political systems, social systems and cultural diversity. Some academic expectations may "define" a smaller portion of a content domain than others, e.g., mathematical structure. (Tables 3–2 through 3–12 are based on data provided by Stuart Kahl at Advanced Systems.)

TABLE 3–2
ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS—GOAL 1
OPEN RESPONSE TESTS

ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
1.1 Accessing Information						
1.2 Reading	✓	✓	✓	✓	✓	✓
1.3 Observing						
1.4 Listening						
1.5–1.9 Math Procedures	✓	✓	✓	✓	✓	✓
1.10 Classifying						
1.11 Writing	✓	✓	✓	✓	✓	✓
1.12 Speaking						
1.13 Visual Arts	✓	✓	✓	✓	✓	✓
1.14 Music	✓	✓	✓	✓	✓	✓
1.15 Movement						
1.16 Technology						

TABLE 3-3
ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS—GOAL 2
OPEN RESPONSE TESTS

ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
2.1 Scientific Skills	✓	✓	✓	✓	✓	✓
2.2 Patterns	✓	✓	✓	✓	✓	✓
2.3 Systems and Interactions	✓	✓	✓	✓	✓	✓
2.4 Models and Scales	✓	✓	✓	✓	✓	✓
2.5 Constancy	✓	✓	✓	✓	✓	✓
2.6 Change Over Time	✓	✓	✓	✓	✓	✓
2.7 Number Concepts	✓	✓	✓	✓	✓	✓
2.8 Mathematical Procedures	✓	✓	✓	✓	✓	✓
2.9 Space and Dimensionality	✓	✓	✓	✓	✓	✓
2.10 Measurement	✓	✓	✓	✓	✓	✓
2.11 Change	✓	✓	✓	✓	✓	✓
2.12 Mathematical Structure	✓	✓	✓	✓	✓	✓
2.13 Data	✓	✓	✓	✓	✓	✓
2.14 Democratic Principals	✓	✓	✓	✓	✓	✓
2.15 Political Systems	✓	✓	✓	✓	✓	✓
2.16 Social Systems	✓	✓	✓	✓	✓	✓
2.17 Cultural Diversity	✓	✓	✓	✓	✓	✓
2.18 Economic Systems	✓	✓	✓	✓	✓	✓
2.19 Geography	✓	✓	✓	✓	✓	✓
2.20 Historical Perspective	✓	✓	✓	✓	✓	✓
2.22 Production	✓	✓	✓	✓	✓	✓
2.23 Analysis of Form	✓	✓	✓	✓	✓	✓
2.24 Aesthetics	✓	✓	✓	✓	✓	✓
2.25 Cultural Heritage	✓	✓	✓	✓	✓	✓

TABLE 3-3 (continued)
ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS—GOAL 2
OPEN RESPONSE TEST

ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
2.26 Cultural Awareness	✓	✓	✓	✓	✓	✓
2.27 Language						
2.28 Second Language						
2.29 Family Life	✓			✓		✓
2.30 Consumerism	✓	✓	✓	✓	✓	✓
2.31 Physical Wellness	✓	✓	✓	✓	✓	✓
2.32 Mental Wellness		✓	✓	✓	✓	✓
2.33 Health Systems	✓			✓	✓	✓
2.34 Psychomotor Skills		✓	✓	✓	✓	✓
2.35 Physical Activities	✓	✓	✓	✓	✓	✓
2.36 Career Path	✓	✓	✓	✓	✓	✓
2.37 Transition Skills	✓	✓	✓	✓	✓	✓
2.38 Post Secondary Search	✓	✓		✓	✓	✓

TABLE 3–4 ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS—GOAL 5 OPEN RESPONSE TESTS						
ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
5.1 Critical Thinking	✓	✓	✓	✓	✓	✓
5.2 Creative Thinking						
5.3 Conceptualizing						
5.4 Decision Making	✓	✓	✓	✓	✓	✓
5.5 Problem Solving	✓	✓	✓	✓	✓	✓

TABLE 3–5 ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS—GOAL 6 OPEN RESPONSE TESTS						
ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
6.1 Multiple Perspectives	✓	✓	✓	✓	✓	✓
6.2 Acquire Knowledge						
6.3 Expanding Knowledge						

TABLE 3–6
ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS—GOAL 1
PERFORMANCE EVENTS

ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
1.1 Accessing Information						
1.2 Reading						
1.3 Observing						
1.4 Listening						
1.5–1.9 Math Procedures			✓		✓	✓
1.10 Classifying						
1.11 Writing						
1.12 Speaking						
1.13 Visual Arts	✓	✓	✓	✓	✓	✓
1.14 Music	✓	✓	✓	✓	✓	✓
1.15 Movement						
1.16 Technology						

TABLE 3-7
ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS—GOAL 2
PERFORMANCE EVENTS

ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
2.1 Scientific Skills	✓	✓	✓	✓	✓	✓
2.2 Patterns	✓	✓			✓	✓
2.3 Systems and Interactions				✓	✓	✓
2.4 Models and Scales	✓	✓	✓	✓	✓	✓
2.5 Constancy	✓				✓	
2.6 Change Over Time						
2.7 Number Concepts				✓		
2.8 Mathematical Procedures	✓				✓	
2.9 Space and Dimensionality	✓	✓		✓		✓
2.10 Measurement	✓		✓			✓
2.11 Change	✓				✓	
2.12 Mathematical Structure			✓			
2.13 Data				✓	✓	✓
2.14 Democratic Principals			✓	✓	✓	✓
2.15 Political Systems		✓	✓	✓		✓
2.16 Social Systems					✓	✓
2.17 Cultural Diversity	✓					
2.18 Economic Systems		✓	✓			
2.19 Geography	✓					✓
2.20 Historical Perspective	✓			✓	✓	✓
2.22 Production	✓	✓	✓	✓		✓
2.23 Analysis of Form	✓	✓	✓	✓	✓	✓
2.24 Aesthetics	✓			✓	✓	✓
2.25 Cultural Heritage	✓				✓	

TABLE 3–7 (continued)
ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS—GOAL 2
PERFORMANCE EVENTS

ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
2.26 Cultural Awareness		✓	✓			
2.27 Language						
2.28 Second Language						
2.29 Family Life	✓					
2.30 Consumerism		✓	✓			
2.31 Physical Wellness				✓		
2.32 Mental Wellness						✓
2.33 Health Systems	✓			✓		✓
2.34 Psychomotor Skills		✓				
2.35 Physical Activities			✓		✓	
2.36 Career Path	✓		✓			
2.37 Transition Skills			✓	✓		
2.38 Post Secondary Search		✓			✓	✓

TABLE 3-8
ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS–GOAL 5
PERFORMANCE EVENTS

ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
5.1 Critical Thinking	✓	✓	✓	✓	✓	✓
5.2 Creative Thinking	✓	✓	✓	✓	✓	✓
5.3 Conceptualizing	✓	✓	✓	✓	✓	✓
5.4 Decision Making	✓	✓	✓	✓	✓	✓
5.5 Problem Solving	✓	✓	✓	✓	✓	✓

TABLE 3-9
ASSESSMENT OF KENTUCKY'S ACADEMIC EXPECTATIONS–GOAL 6
PERFORMANCE EVENTS

ACADEMIC EXPECTATION	GRADE 4		GRADE 8		GRADE 11	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
6.1 Multiple Perspectives	✓	✓	✓	✓	✓	✓
6.2 Acquire Knowledge	✓	✓	✓	✓	✓	✓
6.3 Expanding Knowledge	✓	✓	✓	✓	✓	✓

TABLE 3–10
DISTRIBUTION OF ITEMS ACROSS ACADEMIC EXPECTATIONS
GRADE 4 OPEN-RESPONSE

Content Area	Academic Expectation	1994–95			1995–96		
		Common	Matrix	Total	Common	Matrix	Total
Reading	1.2 Literary	2	13	15	2	13	15
	1.2 Informational	1	6	7	1	5	6
	1.2 Practical/Workplace	1	4	5	1	4	5
	1.2 Persuasive	1	2	3	1	2	3
Science	2.1 Scientific Skills	2	3	5	0	2	2
	2.2 Patterns	3	7	10	3	10	13
	2.3 Systems/Interactions	1	8	9	1	10	11
	2.4 Models/Scales	0	3	3	2	0	2
	2.5 Constancy	1	4	5	0	8	8
	2.6 Evolution	1	7	8	1	8	9
	Inquiry	2	4	6	0	6	6
	Life	1	10	11	4	12	16
	Earth/Space	1	5	6	1	4	5
	Physical	1	10	11	1	4	5
Mathematics	2.7 Number	0	4	4	0	4	4
	2.8 Procedures	2	3	5	1	4	5
	2.9 Space/Dimension	0	4	4	1	3	4
	2.10 Measurement	1	2	3	1	3	4
	2.11 Change	1	2	3	0	3	3
	2.12 Structure	0	4	4	1	2	3
	2.13 Data	1	4	5	1	5	6
Social Studies	2.14 Democratic	0	1	1	0	2	2
	2.15 Political	0	6	6	1	4	5
	2.16 Social	1	3	4	1	5	6
	2.17 Cultural	1	2	3	0	3	3
	2.18 Economic	0	5	5	1	5	6
	2.19 Geography	1	3	4	3	6	9
	2.20 History	2	2	4	1	3	4
		5 Common, 24 Matrix-sampled per area					

TABLE 3–10a
DISTRIBUTION OF ARTS & HUMANITIES AND PRACTICAL LIVING OPEN-RESPONSE ITEMS ACROSS ACADEMIC EXPECTATIONS

Content Area	Academic Expectation	1994–95			1995–96		
		Grade 4	Grade 8	Grade 11	Grade 4	Grade 8	Grade 11
Arts & Humanities	2.22 Production	5	2	1	11	3	2
	2.23 Analysis of Form	4	7	6	10	13	16
	2.24 Aesthetics	7	7	8	9	12	15
	2.25 Cultural Heritage	5	1	1	7	7	7
	2.26 Cultural Awareness	4	8	5	6	10	9
Practical Living/ Vocational Studies	2.29 Family Life	2	0	0	0	2	2
	2.30 Consumerism	1	1	2	2	1	2
	2.31 Physical Wellness	1	2	2	3	2	3
	2.32 Mental Wellness	0	1	1	3	3	1
	2.33 Health Systems	1	0	1	1	2	2
	2.34 Psychomotor Skills	0	1	0	1	1	1
	2.35 Physical Activities	2	1	4	3	1	4
	2.36 Career Path	2	5	1	4	4	3
	2.37 Transition Skills	2	1	3	6	2	6
	2.38 Post Secondary Search	2	1	2	3	5	4
		12 Matrix-sampled per area			24 Matrix-sampled per area		

TABLE 3–11
DISTRIBUTION OF ITEMS ACROSS ACADEMIC EXPECTATIONS
GRADE 8 OPEN-RESPONSE

Content Area	Academic Expectation	1994–95			1995–96		
		Common	Matrix	Total	Common	Matrix	Total
Reading	1.2 Literary	2	9	11	2	10	12
	1.2 Informational	2	9	11	2	10	12
	1.2 Practical/Workplace	1	4	5	1	3	4
	1.2 Persuasive	1	2	3	1	2	3
Science	2.1 Scientific Skills	1	6	7	3	6	9
	2.2 Patterns	3	8	11	2	6	8
	2.3 Systems/Interactions	5	14	19	3	12	15
	2.4 Models/Scales	3	10	13	2	6	8
	2.5 Constancy	2	8	10	2	5	7
	2.6 Evolution	2	10	12	3	7	10
	Inquiry	2	5	7	2	5	7
	Life	2	7	9	1	9	10
	Earth/Space	1	5	6	1	6	7
	Physical	1	7	8	2	6	8
Mathematics	2.7 Number	0	4	4	1	5	6
	2.8 Procedures	1	6	7	2	6	8
	2.9 Space/Dimension	2	3	5	1	3	4
	2.10 Measurement	1	6	7	2	4	6
	2.11 Change	1	6	7	2	4	6
	2.12 Structure	1	2	3	0	2	2
	2.13 Data	1	3	4	2	3	5
Social Studies	2.14 Democratic	2	2	4	1	4	5
	2.15 Political	1	3	4	0	3	3
	2.16 Social	1	3	4	1	1	2
	2.17 Cultural	1	5	6	1	3	4
	2.18 Economic	1	2	3	1	3	4
	2.19 Geography	2	5	7	1	5	6
	2.20 History	0	5	5	1	5	6
		6 Common, 24 Matrix-sampled per area					

TABLE 3–12
DISTRIBUTION OF ITEMS ACROSS ACADEMIC EXPECTATIONS
GRADE 11 OPEN-RESPONSE

Content Area	Academic Expectation	1994–95			1995–96		
		Common	Matrix	Total	Common	Matrix	Total
Reading	1.2 Literary	2	7	9	2	8	10
	1.2 Informational	2	9	11	2	6	8
	1.2 Practical/Workplace	1	5	6	1	5	6
	1.2 Persuasive	1	5	6	1	5	6
Science	2.1 Scientific Skills	3	13	16	2	14	16
	2.2 Patterns	4	20	24	5	19	24
	2.3 Systems/Interactions	6	16	22	5	19	24
	2.4 Models/Scales	1	11	12	3	7	10
	2.5 Constancy	3	12	15	4	13	17
	2.6 Evolution	3	9	12	2	13	15
	Inquiry	2	7	9	1	7	8
	Life	3	8	11	1	10	11
	Earth/Space	2	6	8	3	4	7
	Physical	2	6	8	2	6	8
Mathematics	2.7 Number	0	2	2	1	3	4
	2.8 Procedures	2	4	6	0	5	5
	2.9 Space/Dimension	0	5	5	1	6	7
	2.10 Measurement	2	4	6	1	5	6
	2.11 Change	1	4	5	1	5	6
	2.12 Structure	1	2	3	1	2	3
	2.13 Data	1	5	6	1	7	8
Social Studies	2.14 Democratic	1	4	5	0	5	5
	2.15 Political	1	4	5	3	4	7
	2.16 Social	1	7	8	2	6	8
	2.17 Cultural	2	2	4	1	4	5
	2.18 Economic	1	4	5	1	7	8
	2.19 Geography	2	4	6	1	6	7
	2.20 History	2	10	12	2	9	11
		6 Common, 24 Matrix-sampled per area					

CURRICULUM FRAMEWORK

The academic expectations describe what students are expected to do with the knowledge and skills they acquire in school. They do not define in detail all the knowledge and individual skills students need. The Kentucky Education Reform Act charged the Department of Education to develop a model curriculum framework that could be used by teachers and school administrators in designing local curriculum.

More than 100 teachers, counselors, administrators, regional service center consultants, and university personnel were significantly involved in the development of *Transformations: Kentucky's Curriculum Framework*. Volume I of the curriculum framework was designed to help administrators and teachers address the academic expectations in their local or school curricula. Each academic expectation is defined by a set of demonstrators which are benchmarks or indicators of student progress toward the expectation. The demonstrators are accompanied by learning links and related concepts which provide teachers with ideas about how to incorporate the academic expectations in their curricula. Also included are sample activities, which suggest ways to involve students in engaging instructional/assessment experiences. The activities are coordinated with the demonstrators and the statewide assessment.

The publication of *Transformations* provided Kentucky's educators with a document that integrates the processes of curriculum, instruction, and assessment. It was an important step in helping join the power of the assessment with positive instruction.

CONTENT GUIDELINES AND CORE CONTENT

The 57 academic expectations are broad and comprehensive. The content for the assessment was further specified beginning in 1994. Content Guidelines were used in developing and selecting items for the 1994-95 and 1995-96 open-response tests. Core Content for Assessment, which is essentially a subset of Content Guidelines, was referenced in selecting 1995-96 test items, and implemented fully in the development and selection of items for 1996-97 testing.

KIRIS NAEP-BASED FRAMEWORKS

The content domains for the major subject areas tested are described by the assessment frameworks, which detail lists of subdomains and cognitive processes within each subject area. The frameworks, based on those employed in creating NAEP tests, serve three purposes. First, they reflect an underlying rationale for the testing approach, based on current theory or research. Second, they guide the process of test development, helping to ensure that coverage of a subject area domain is balanced and complete. Third, they provide a definition of reporting categories—subdomains for which test results are produced and reported. In the following sections, the frameworks used for multiple-choice and open-response questions at accountability grades and for the voluntary continuous assessment tests (in the non-accountability grades) are presented in detail for reading, mathematics, science, social studies, writing, the arts and humanities, practical living, and vocational studies.

READING. The reading assessment framework is based on the principle that reading is an interaction among the reader, the text, and the situation. The reader interacts with the text, using prior knowledge and a repertoire of reading strategies gained through experience to make sense of and extend meaning from the text for a purpose. Kentucky's reading framework, similar to that of the 1992 NAEP framework, is based on this definition of reading and on Kentucky's academic expectation that all students will exhibit basic communication skills. The categories are described in Table 3–13. For each testing year in the accountability cycle, the set of test forms covers each indicated area in the table.

TABLE 3–13
KIRIS READING CONTENT FRAMEWORK

Item Category	Passage Type			
	Literary	Informa-tional	Practical/ Workplace	Persuasive
Comprehension: Making Sense of What was Read	✓	✓	✓	✓
Comprehension: Elaborating and Responding Critically	✓	✓	✓	✓
Reader Awareness				
Reader Response				

PASSAGE TYPE. The passages used in the assessment are drawn from authentic texts, rather than those written solely for the purpose of testing. The passages appeared with pictures and headings as originally published, when appropriate or available. The quality of the writing and age-appropriateness of the material covered were additional selection criteria.

Effective readers approach different kinds of materials in different ways based on the purposes in reading the type of text, thus passages were chosen from:

- literary pieces, such as excerpts from novels, short stories, poems, plays, letters, or essays;
- informational pieces, such as articles from newspapers or magazines, content area textbook sections, or encyclopedias;
- practical/workplace pieces, such as transportation schedules, directions, or applications; and
- persuasive pieces, such as speeches, editorials, or advertisements.

Literary, informational, and practical passages are often associated with different purposes—reading for literary experience, reading to be informed, or reading to perform a task. The fourth passage type (persuasive) includes literary, informational, and practical passages whose purpose is largely to influence the opinion of readers.

Table 3–14 shows the approximate percentages of passage types used at each level for the open-response tests that were administered each year during the second accountability cycle.

TABLE 3–14
PASSAGE TYPES USED IN THE KIRIS READING ASSESSMENT

Grade	Passage Type			
	Literary	Informational	Practical/ Workplace	Persuasive
4	51%	22%	17%	10%
8	38%	37%	15%	10%
11	30%	30%	20%	20%

ITEM CATEGORY. Effective readers demonstrate the ability to understand information presented in the text and to analyze and extend information based on their own experiences and prior knowledge. The framework for the Kentucky assessment reflects the belief that these behaviors should be practiced by readers at every level of development and should not be viewed as developing sequentially.

Making Sense of What was Read requires that readers formulate an initial, general impression of the material and continue to interpret the text as they read. The reader focuses on important portions of the text and links information from different sections by examination and analysis. The following questions and statement are typical of this process of making sense of what was read:

- What is the story about?
- Describe the main character.
- How did the character change from beginning to end?
- What is the article about?
- What caused this particular event?
- What is the author's position?
- What support does the author present for his/her position?

Elaborating and Responding Critically demands that readers connect their own knowledge with text and evaluate the text. Readers also extend, examine, analyze, apply, and judge the ideas in the text. The following questions and statement demonstrate this category:

- How does this story compare/contrast with your own experiences?
- Would you recommend this story? Why or why not?
- Does this remind you of a current event? Explain.
- Compare this story/article to that one.
- What additional information would you need to complete this task?
- What would happen if you omitted step X?
- Do you agree with the author's position? Why or why not?

Reader Awareness. Effective readers demonstrate the ability to monitor and regulate their comprehension of text. Questions in this category address awareness/comprehension of reading strategies, such as:

- setting purposes for reading;
- previewing the text;
- identifying the type (genre) of the text;
- activating prior knowledge about the topic;
- visualizing while reading;
- using text structure and cues (e.g., italics);
- making predictions;
- paraphrasing and summarizing;
- seeking clarification;
- raising questions while reading;
- reading ahead or rereading;
- taking notes;
- reviewing comprehension.

Reader awareness questions were asked only as multiple-choice pre-test items and did not count towards school accountability scores.

Reader Attitude. Effective readers value reading and are likely to read frequently, which supports continued improvement of their reading skills. This component of the assessment examined students' perceptions about reading and about themselves as readers. Reader attitude questions were asked only as multiple-choice pre-tests items and did not count toward school accountability scores.

MATHEMATICS OPEN-RESPONSE. Table 3-15 shows the content framework used to build the mathematics tests. Because of the interconnected nature of mathematics (and the richness of KIRIS open-response items), a single item often addresses more than one content and process category.

TABLE 3-15
KIRIS MATHEMATICS CONTENT FRAMEWORK

Content	Process			
	Conceptual Knowledge	Procedural Knowledge	Problem Solving	Reasoning & Analysis
Number	✓	✓	✓	✓
Mathematical Procedures	✓	✓	✓	✓
Space & Dimensionality	✓	✓	✓	✓
Measurement	✓	✓	✓	✓
Change	✓	✓	✓	✓
Mathematical Structure	✓	✓	✓	✓
Data	✓	✓	✓	✓

The content categories are the academic expectations for mathematics listed under "Goal 2—Core Concepts and Principles" by Kentucky's Council on School Performance Standards. To clarify the mathematics content categories, Table 3–16 presents an overview of mathematics topics that belong to each one. The process categories are from the *Curriculum and Evaluation Standards for School Mathematics* developed by the National Council of Teachers of Mathematics (NCTM). The entire framework is consistent with the most recent NAEP framework for mathematics. The process categories, consistent with those described in the NCTM *Standards*, represent different thought processes or ways of mentally processing information. They are not mutually exclusive.

TABLE 3–16
TOPICS ADDRESSED IN EACH MATHEMATICS CONTENT CATEGORY

Number	Mathematical Procedures	Space and Dimensionality
Number Relations Number Sense Infinity Notation Systems Estimation Number Theory Equivalent Forms Percents	Arithmetic Proportion Computational Algorithms Algebraic Processes	Regular and Irregular Figures in Various Dimensions Spatial Visualization Congruence Symmetry Transformations Similarity Coordinate Systems Vectors Orientation
Measurement	**Change**	**Mathematical Structure**
Direct and Indirect Measurements Maximum, Minimum Metric, Customary, and Nonstandard Units Length, Area, Volume, Capacity Mass, Weight Time, Rate Temperature Trigonometry Angle Curves Money	Transformations Algebraic, Matrix, and Trigonometric Representations Limit Series Continuous vs. Discrete Sequences Curve Fitting Functions Patterns Variables	Proof Fields Axiomatic Systems Logic Number Systems Order Relations Groups Discrete Structures Order of Operations
Data		
Applied Probability Theoretical Probability Probability in Experiments Measures of Central Tendency (including mean, median, mode) Measures of Dispersion (including range and outliers) Statistical Procedures and Decisions		

Test questions assessing students' conceptual knowledge in mathematics provide evidence of their abilities to label, verbalize, and define the concept; recognize and generate examples and non-examples; use models, diagrams, and symbols to represent the concept; translate from one mode of representation to another; and compare, contrast, and integrate related concepts.

Procedural knowledge requires students to demonstrate their abilities to reliably and efficiently execute procedures and verify their results empirically (e.g., using models) or analytically; verbalize and give reasons for the steps in a procedure; recognize correct and incorrect procedures; generate new procedures and extend or modify familiar ones; and recognize situations in which a procedure is appropriate, necessary, and correctly applied. Procedures include, but are not limited to, computational algorithms.

Students' ability to use their knowledge of mathematics for problem solving is measured by questions that require that they recognize and formulate problems by identifying the key features and questions in a given situation; solve problems by selecting and using appropriate strategies, models, and relevant mathematical content; and verify and generalize solutions by reflecting critically on the problem.

The assessment of mathematical reasoning and analysis requires students to use standard forms of reasoning such as inductive, deductive, and statistical inference; use contextual reasoning such as proportional and spatial reasoning; and identify and examine components of a given situation to abstract commonalities, specify properties and relationships, and identify underlying structure.

MATHEMATICS PORTFOLIOS. For portfolios, students select their best examples of mathematics within the content requirements set by the Kentucky Mathematics Advisory Committee. Those content requirements ensure that portfolios allow students to demonstrate that they meet the requirement of the pertinent academic expectations, as defined by the 7 Core Concepts from *Transformation: Kentucky's Curriculum Framework* and the 4 process skills for the National Council of Teacher's of Mathematics, *Curriculum and Evaluation Standards*. Table 3–16a presents the entries required for each accountability grade level during Cycle 2.

SOCIAL STUDIES. The development and selection of test questions for the social studies component of the assessment were guided by the two-dimensional matrix shown in Table 3–17.

The major content categories represent subdisciplines of social studies. The seven content categories listed are the social studies academic expectations identified by Kentucky's Council on School Performance Standards. The cognitive skills represent types of mental or cognitive processing consistent with Bloom's Taxonomy of Cognitive Objectives. These skills refer to one dimension of a question's characteristics, and the content categories refer to another. For example, a question requiring "higher order thinking skills" as well as an understanding of content would be placed in the appropriate content category (Democratic Principles, Cultural Diversity, etc.) and be identified as "higher order" according to the cognitive skill dimension. Social studies achievement results were grouped into broader categories for reporting.

TABLE 3–16a
KIRIS MATHEMATICS PORTFOLIO CONTENT REQUIREMENTS

Quality Criteria (All Grades)	Portfolio Pieces (All Grades)
Problem Solving Understand methods for solving problems and extend the problem **Reasoning** Think by using mathematical ideas and verify your solution by using logical explanations **Mathematical Communication** explain mathematics to others using a variety of representations **Understand/Connecting Core Concepts**	1. **Table of Contents** - a listing of the titles of the pieces in your portfolio. 2. **Student Signature Sheet** - a statement you will sign indicating that the work in the portfolio is you own original work. 3. **Letter to the Reviewer** - a letter you write to the person who will read your portfolio. In it you might write about what you've learned from keeping a Mathematics Portfolio; explain which entry is you best and why; tell from which entry you learned the most and why; explain what pleased you about the entries in your portfolio; tell what you observed in reviewing your entries. 4. **5-7 Best Pieces** - pieces that demonstrate a variety of ways to use mathematics; involve all core concepts; number, mathematical procedures, space and dimensionality, measurement, change, mathematical structure, data; and use different mathematical tools. a. investigations - using inductive reasoning to search for answers to questions b. applications- using mathematical concepts, principles, and procedures to solve real world problems c. non-routine - combining or inventing problem solving strategies to arrive at solutions or results d. projects - developing extended mathematical activities e. interdisciplinary - using mathematics meaningfully with other subjects f. writing- writing about mathematics to explain your thinking and/or solutions

TABLE 3– 17
KIRIS SOCIAL STUDIES CONTENT FRAMEWORK

Content	Cognitive Skills	
	Knowledge and Comprehension	Application and Higher Order
Democratic Principles	✓	✓
Structure and Function of Political Systems	✓	✓
Structure and Function of Social Systems	✓	✓
Cultural Diversity	✓	✓
Structure and Function of Economic Systems	✓	✓
Relationship of Geography to Human Activity	✓	✓
Historical Perspective	✓	✓

BEST COPY AVAILABLE

The cognitive skills represent a hierarchy of mental processes required of students in answering the test questions. "Knowledge" questions require primarily memory processes addressing lower-level skills such as factual recall. "Comprehension" questions assess the understanding of concepts, and generally require students to demonstrate understanding by translating or explaining (or, in the case of multiple-choice questions, identifying appropriate translations or explanations). "Application" questions assess a slightly higher level of understanding by requiring students to apply their understanding of concepts to particular situations. Bloom's "higher order" processes include analysis, synthesis, and evaluation. Analysis questions require that information be broken down into parts so that ideas and their interrelationships are made explicit. Synthesis questions require that ideas be pulled together to produce a broader message or entity. Evaluation questions require students to make judgments about the relative value of ideas and approaches. In pre-test multiple-choice questions assessing this process, several response options often have some merit, but one is clearly "best" for any particular question.

SCIENCE. The conceptual framework used to guide the development of the science component of the assessment is shown in Table 3–18.

TABLE 3–18
KIRIS SCIENCE CONTENT FRAMEWORK

Science Content	Themes of Science					
	Nature of Scientific Activity	Constancy	Evolution	Models and Scales	Patterns	Systems and Interactions
Scientific Inquiry Nature of Science/Investigations Data Gathering/Reduction Data Analysis/Interpretation	✓	✓	✓	✓	✓	✓
Life Science Characteristics of Life Organisms Ecology/Environment	✓	✓	✓	✓	✓	✓
Earth/Space Science Astronomy Meteorology Geology/Natural Resources	✓	✓	✓	✓	✓	✓
Physical Science Matter Energy Force/Motion	✓	✓	✓	✓	✓	✓

The two major crossing dimensions, *Science Content* and *Themes of Science* include components of the 1994 NAEP science framework. *Scientific Inquiry* was added to the science content areas to accommodate questions on scientific inquiry which require little or very low levels of specific science content knowledge. This allowed for the reporting of results to reveal program strengths or weaknesses with respect to inquiry skills. Other questions involving inquiry skills, but which required a greater degree of content understanding as well, were included in their appropriate fields of science.

The themes of science are the science academic expectation categories of the Council on School Performance Standards. These themes cross the disciplines of science and are intended to help unify instruction and thinking across those fields. Because these themes are far from mutually exclusive (that is, a single question was likely to address several themes simultaneously), they are treated in the Kentucky framework just as they are by NAEP. Questions are counted under more than one theme for the reporting of results on the Council's academic expectations.

None of the dimensions so far described cleanly separate the questions by cognitive process levels. Therefore, questions were also identified as either *Knowledge/Comprehension* or *Application/Higher Order*. This assured that the questions in a particular category of another dimension were not overly dominated by either lower or higher order skills.

WRITING. Proficiency in writing is measured using portfolios and an on-demand writing exercise (although for the first two accountability cycles, only the Writing Portfolio counted toward accountability decisions).

PORTFOLIOS. For portfolios, students select their best examples of writing within the content requirements set by the Kentucky Writing Advisory Committee. Those content requirements ensure that portfolios allow students to demonstrate that they meet the requirement of the pertinent academic expectation, ***Students demonstrate the ability to communicate through their writing ideas and information to a variety of audiences for a variety of purposes in a variety of written forms.*** Table 3–19 presents the entries required for each accountability grade level during Cycle 2.

TABLE 3–19
KIRIS WRITING CONTENT REQUIREMENTS

Quality Criteria (All Grades)	Portfolio Pieces		
	Grade 4	Grade 8	Grade 12
Correctness (spelling, punctuation, capitalization) **Language** (usage, word choice) **Sentences** (correctness, structure, effectiveness) **Organization** (logic, coherence, transitions) **Idea Development and Support** **Purpose and Audience**	1. Table of Contents 2. A Letter to the Reviewer 3. One personal narrative 4. One short story, poem, play/script, or piece of fiction 5. One piece of writing the purpose of which is to: a. present/support a position, idea, or opinion; or b. tell about a problem and its solution; or c. inform 6. One piece of writing from a study area other than English/ Language Arts. 7. A Personal Selection	1. Table of Contents 2. A Letter to the Reviewer 3. One personal narrative 4-5. Two pieces of writing, each of which will achieve any one or more of the following purposes: a. predict an outcome b. defend a position c. analyze or evaluate a situation, person, place or thing d. solve a problem e. explain a process or concept f. draw a conclusion g. create a model 6. One short story, poem, play/script, or other piece of original fiction 7. A Personal Selection A minimum of one piece of writing must come from a subject area other than English or Language Arts.	1. Table of Contents 2. A Letter to the Reviewer 3. One personal narrative, memoir, personal essay, or personal vignette 4. One short story, poem, or play/script 5-7. Three pieces of writing, each of which will achieve any one or more of the following purposes: a. predict an outcome b. defend a position c. analyze or evaluate a situation, person, place or thing d. solve a problem e. explain a process or concept f. draw a conclusion g. create a model A minimum of two pieces of writing must come from a subject area other than English or Language Arts.

BEST COPY AVAILABLE

ON-DEMAND. During Cycle 2 a total of twenty-four prompts were administered on twelve forms. Student responses were scored using the same scoring guide developed for writing portfolios, as described in Chapter 4.

Students chose to respond to one of the two prompts. Using a choice from two prompts made the on-demand writing situation more closely resemble an authentic writing process. General directions encouraged prewriting or planning and provided editing and revising suggestions that paralleled the criteria by which the writing was evaluated. Prompts were chosen to be consistent with the amount of time students were able to sustain their writing: ninety minutes.

ARTS AND HUMANITIES. The framework for the assessment of arts and humanities is based on Kentucky's core concepts and related academic expectations. Table 3–20 presents the content framework.

TABLE 3–20
KIRIS ARTS AND HUMANITIES CONTENT FRAMEWORK

Academic Expectations	Area				
	Visual Arts	Music	Dance	Theatre	Humanities
Create products and make presentations that convey concepts and feelings	✓	✓		✓	
Analyze artistic products and performances	✓	✓	✓	✓	✓
Appreciate creativity and the values of the arts and the humanities	✓	✓	✓	✓	✓
Show an understanding of the influence of time, place, personality, and society on the arts and humanities	✓	✓	✓	✓	✓
Recognize differences and commonalities in the human experience	✓	✓	✓	✓	✓
Demonstrate awareness of the diversity of forms, structures, and concepts across languages and how they interrelate.	Not Assessed				
Understand and communicate in a second language.	Not Assessed				

The arts and humanities framework for all grade levels focuses on five major areas: the visual arts, music, dance, theatre, and Humanities. Within those areas, the Arts and Humanities Advisory Committee identified specific knowledge and skills to be assessed. Table 3–21 presents additional information about the content of the Arts and Humanities test.

TABLE 3–21
ADDITIONAL INFORMATION ABOUT THE CONTENT OF THE ARTS AND HUMANITIES TEST

Visual Arts	Music	Dance	Theatre
1. Basic understanding of elements: Line Shape Form Texture Space Value Color 2. Six basic ways of making art: Drawing Painting Print Ceramics Sculpture Textiles 3. Elements and principles of design 4. Drawing from real life 5. Analysis 6. Appreciation 7. How do visual arts communicate meaning? 8. Historical and cultural awareness	1. Basic understanding of elements: Rhythm Melody Harmony Form Expression of Style Texture Tonality 2. Know basic notations 3. Critique a work of past or present based on basic understanding 4. How is meaning communicated through music? 5. Historical and cultural awareness	1. Elements of dance: Movement Space Time Force Technique Expression 2. Analysis 3. Appreciation—ability to communicate your own value system 4. How does dance communicate aesthetic value? (How is meaning portrayed through movement?) 5. Historical and cultural awareness	1. Express through movement and expressions 2. Analyze a theatre production (interpret) 3. How to role-play 4. Basic understanding of dramatic elements: Character Plot Theme Language Sound Spectacle 5. Productive elements: Scenery Costumes Props Lighting 6. Theatre Careers: Actor Director Playwright Producer Designers (light and scenery, make-up, costume) 7. Historical and cultural awareness

PRACTICAL LIVING AND VOCATIONAL STUDIES. The assessment frameworks developed for practical living and vocational studies are somewhat different from those used in the traditional academic areas but similar to the framework developed for the arts and humanities.

PRACTICAL LIVING. The practical living assessment addresses the development of health and fitness habits and behaviors; child and family development; consumer and resource management; and other concepts that support and will effect students' growth, development, and wellness throughout their lives. It encompasses the traditional areas of health education, home economics, and physical education, but promotes both instruction and assessment that integrate those disciplines throughout the school years.

The practical living assessment is developed using core concepts based directly on pertinent academic expectations. These core concepts are presented in Table 3–22.

VOCATIONAL STUDIES. Vocational studies emphasizes the skills that all students need to make the successful transition from school to the world of work and to sustain careers that will enrich their lives.

The focus of the vocational studies assessment is further defined by student education level. At the elementary level, students are evaluated on their developing appreciation of work, their awareness of themselves as workers, and their developing knowledge of career options. In the middle grades, the assessment focuses on students' investigations of career options, their knowledge of the relationships between careers and life roles, and their abilities to connect educational achievement to career opportunities. By the time they complete high school, students are expected to know the demands of a career and how it will affect their multiple roles in life. Additionally, students are to have acquired the knowledge and skills necessary for making a successful transition to the path they have selected for the immediate future—college, technical school, military service, or other work.

The core concepts and academic expectations addressed in the vocational studies assessment are also presented in Table 3–22.

TABLE 3–22
KIRIS PRACTICAL LIVING/VOCATIONAL STUDIES CONTENT FRAMEWORK

Practical Living	Vocational Studies
Individual Well-Being and Healthy Family Relationships Consumer Skills Physical Wellness Mental and Emotional Wellness Community Service/Resource Systems Physical Movement Skills Lifetime Physical Activities	Choosing and Preparing for a Career Skills and Work Habits for Success in Future Schooling and Work Skills needed to be Accepted into College, Post-secondary Training, or job

ALTERNATE PORTFOLIOS. Students who are severely handicapped such that they are not expected to obtain a high school diploma are eligible to develop an Alternate Portfolio rather than participate in other aspects of the KIRIS program. Such students represent about one-half of one percent of Kentucky's school population. Specific criteria for eligibility are detailed in KDE Program Advisory 92-OCAA-104 (February, 1993) and published in the *Kentucky Alternate Portfolio Teacher's Guide* and the annual *District Assessment Coordinator Implementation Guidebook*.

The Alternate Portfolio targets 28 Academic Expectations selected by the Alternate Portfolio Committee as being most assessable with a portfolio. A complete alternate portfolio includes:

- Table of Contents completed according to provided guidelines;
- Letter to the reviewer, describing the portfolio and its contents, written or dictated by the student (or a collaboration between the student and a non-disabled peer);
- Seven to ten entries representing the breadth of types, contexts, and domain areas;
    - One entry must be the student's weekly schedule and a description of its use, indicating the types of activities included, and the opportunities afforded for choice and for interaction with peers without disabilities;

  - At eighth and twelfth grades, one entry must be a résumé of volunteer and paid job experiences, accompanied by employer evaluations;
  - Photographs, audio- or videotapes, computer disks, and the like must be accompanied by a paragraph describing the entry and the rationale for its inclusion;
  - Student program data must be graphed and accompanied by a copy of the instructional program and a qualitative response from the student, non-disabled peer, and/or family member or caregiver;
  - One entry may be a sample of the student's present mode(s) of communication and description of its use;
- Letter from a family member or caregiver validating the contents of the portfolio.

ENTRY TYPES. Alternate portfolios include writing, projects, instructional program data, applications, and investigations or discoveries. Individual entries may span two or more types.

CONTEXT. Alternative Portfolio entries should demonstrate integration, functionality, and age-appropriateness. They may use computer, assistive/adaptive, and other technology, and provide opportunities for choice.

DOMAIN. Appropriate domain areas include vocational, recreation or leisure, and personal management activities.

GROUP ENTRIES. The student may produce entries alone or as part of a group. Group entries must include an explanation and a reflection by the individual student.

CONTENT GUIDELINES AND CORE CONTENT FOR ASSESSMENT

As the results of the first KIRIS assessment were being analyzed, it became apparent that educators across the Commonwealth desired additional information regarding the focus of the test items. This led to the development of detailed content guidelines to be used both to guide the development of future test forms and to provide teachers with guidance about targeting their instruction. Content Guidelines were used in developing and selecting items for the 1994-95 and 1995-96 open-response tests. Core Content for Assessment, which is essentially a subset of Content Guidelines, was referenced in selecting 1995-96 test items, and implemented fully in the development and selection of items for 1996-97 testing. Copies of the Content Guidelines and Core Content for Assessment are available from KDE.

ITEM TYPES

KERA mandated that the assessment program be "primarily performance-based." In addition to open-response item types, KIRIS used portfolios and performance events during Cycle 2. Performance events were administered but were not included for accountability purposes during Cycle 2.

PORTFOLIOS. Kentucky writing portfolios are compilations of five to seven best pieces of student work per portfolio. For 1995–96, all grade 4, 8, and 12 students enrolled in a Kentucky public school were required to complete a writing portfolio (with minor exceptions as stipulated by the Kentucky Department of Education). Pieces for inclusion in these portfolios were produced throughout the school year as a part of the students' classroom curriculum.

The writing portfolio has specific requirements for each grade level that include a variety of types of work and content area requirements. The writing portfolio also requires a Table of Contents, a Letter to the Reviewer reflecting on the student's work, and a Student Signature Sheet verifying the student's authorship.

The mathematics portfolio also has specific requirements for each grade level that include a variety of types of work and content area requirements. The mathematics portfolio requires a Table of Contents, a Letter to the Reviewer reflecting on the student's work, and a Student Signature Sheet verifying the student's authorship.

In the fall of each year, students were provided with Student Brochures that contained information about requirements, standards, and development issues. Teachers received teacher handbooks and training to assist them in working with students to develop portfolios. When portfolios were completed, they were scored by Kentucky classroom teachers. To prepare for scoring, teachers received the Holistic Scoring Guides, benchmark portfolios, and training portfolios. In addition, teachers received training in standards and scoring procedures.

In 1992–93, an Alternate Portfolio was instituted for students with severe handicapping conditions likely to prevent them from ever obtaining a high school diploma. The Alternate Portfolio is described above.

PERFORMANCE EVENTS. In the second accountability cycle, performance events in mathematics, science, social studies, arts and humanities, and practical living/vocational studies were administered for grades 4, 8, and 11 (but at grades 4, 8, and 12 for 1992-93 and 1993-94). Although administered, performance events were not included in accountability calculations. A typical performance event included a period of group work on a problem in a particular subject, followed by individual written work on the same problem. There were two tasks in each content area of mathematics, science, and social studies plus three arts and humanities, two practical living, and two vocational studies tasks. The events were administered by performance events facilitators who brought materials to the schools, set them up, administered the events, and returned the response forms to Advanced Systems for scoring. In all but a few very small schools, each student took just one event in a group with a few other students. Schools with fewer than nine students were exempt from testing. Figure 3–1 presents an example of a performance event item. This Item, from Grade 11 Mathematics, is typically presented as a sealed booklet. Instructions have been simplified to allow for presentation in this Technical Manual.

OPEN-RESPONSE QUESTIONS. Open-response questions were administered in reading, mathematics, science, social studies, arts and humanities, and practical living/vocational studies. Each open-response question is written so that students should be able to complete their responses in about ten minutes. Figure 3–2 presents an example of an open-response item with a scoring guide. Figure 3–3 gives an annotated student response to that same open-response question.

Open-response questions were included because they produce reliable scores, and they are reflective of KERA's educational direction. First, they require students to generate answers rather than merely recall or recognize information. Second, they emphasize student communication skills, and thus encourage teachers to use more powerful instructional practices that utilize communication skills as well as other aspects of problem solving.

SUMMARY OF ITEM TYPES ADMINISTERED EACH YEAR. Changes over time include the emphases given to the different types of testing, the way testing is conducted, and the content areas to be included in each year. Tables 3–23 through 3–25 provide a general description of the tests administered at grades 4, 8, and 11 in 1994–95 and 1995–96.

FIGURE 3–1
EXAMPLE OF A PERFORMANCE EVENT ITEM

MINIATURE GOLF

GENERAL INSTRUCTIONS

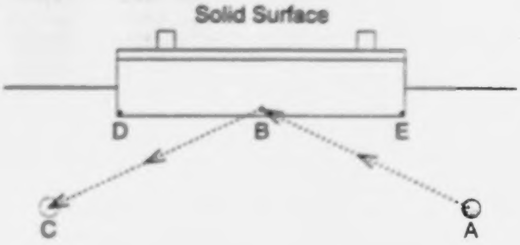
Write your name, school name, and school code on the lines above. **Do not open this packet until the directions tell you to do so.** You will have a total of 45 minutes to complete this task. You may use up to 20 minutes to complete the group work, found on the front and back pages of this form.

Your group should have the following materials:

- rubber ball, string, masking tape, and chalkboard protractor for group task
- pencil, ruler, protractor, and scratch paper for each student
- 1 bag of pattern pieces per student (orange, blue, red, and green)
- wood "bank"

GROUP ACTIVITY

In this task your group will investigate how a ball rebounds when "banked" against a surface. You will conduct an experiment and use the results.



	m∠EBA	m∠CBD
1st trial		
2nd trial		
3rd trial		

Select students from the group to perform the tasks detailed on the right.

Student #1, standing at some point A with a rubber ball, rolls the ball along the floor toward the wood "bank" as shown in the diagram at the left. **It is important to roll the ball with only enough force to cause it to bank in a straight line.**

Student #2 notes the place the ball strikes the bank and marks it with a piece of tape (point B).

Student #3 stops the ball at some point C along the path of the deflected ball.

Now, one student should stretch a piece of string straight between points A and B and tape the string in place. Next, someone should stretch a piece of string straight along the path from B to C and tape that string in place. Use a chalkboard protractor to measure ∠ABE and ∠CBD. Each student must record the results in the table above.

FIGURE 3–1 (continued)

1. ∠ EBA is called the *angle of incidence* of the ball's path to the bank. ∠ CBD is called the *angle of reflection* of the ball's path away from the bank.

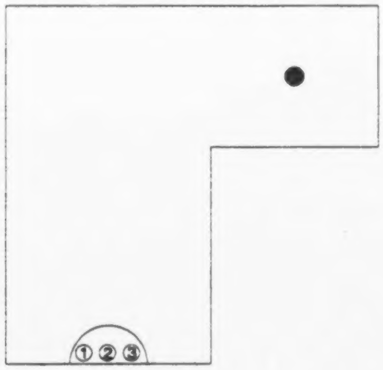
   Theoretically, the angle of incidence equals the angle of reflection. If your group's angles of incidence and reflection were not equal, what factors could have contributed to the inaccuracy of your measurements?

____

____

____

2. The statement above, that the angle of incidence equals the angle of reflection, is put to use when playing billiards or miniature golf. The diagram below shows one hole at a miniature golf course.



The golf ball may be placed at position 1, 2, or 3 at the "tee off" or starting position. In order for the ball to go directly into the hole in one shot (a hole-in-one), the ball must be hit so that it banks off of at least one of the walls of the course boundary.

Discuss with your group the possible paths the ball could take when started from one of the tee-off positions and banked off of a wall for a hole-in-one.

Draw a path for a hole-in-one on the diagram at the left. Label the measure of the angle of incidence and the measure of the angle of reflection.

FIGURE 3–2

EXAMPLE OF AN OPEN-RESPONSE ITEM, WITH SCORING GUIDE

Twenty people took part in a basketball free throw competition. The first shooter scored 20 points, the second shooter scored 28 points, the third shooter scored the average of the number of points scored by the first two shooters, and each of the next shooters scored the average of all the previous shooters.

a. How many points did the fourth shooter score?

b. How many points did the sixth shooter score?

c. If the first shooter scored **X** points and the second shooter scored **Y** points, how many did the 20th shooter score?

d. How would the results in this problem be changed if the fourth shooter were the first to score the average of all the previous shooters? Explain your reasoning.

BE SURE TO LABEL YOUR RESPONSES (a), (b), (c) AND (d).

SCORING GUIDE

Score	Description
4	Student gives correct solutions and strategy for finding each shooter's score. Response shows ability to generalize. Discussion is appropriate, showing an understanding of averages for more than two scores.

Correct answers:

a. $\frac{20 + 28}{2} = 24$ $\frac{20 + 28 + 24}{3} = \frac{72}{3} = 24$

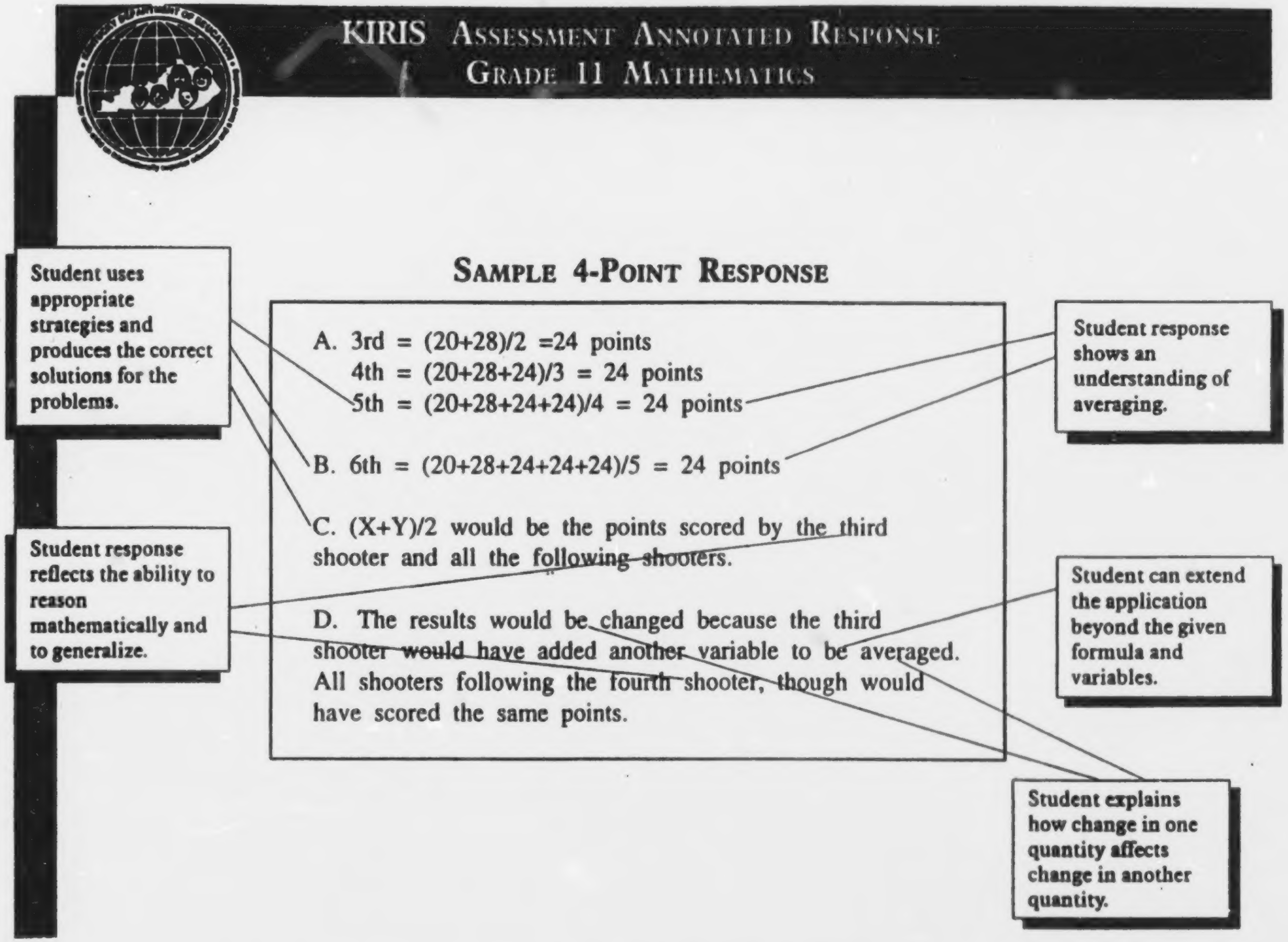
b. 6th
24

c. 20th shooter scored $\frac{x + y}{2}$

d. Each shooter starting with the fourth would score the average of the first three shooters.
$\frac{x + y + z}{3}$

BEST COPY AVAILABLE

FIGURE 3–3
ANNOTATED RESPONSE TO OPEN-RESPONSE QUESTION



KIRIS Assessment Annotated Response
Grade 11 Mathematics

Sample 4-Point Response

A. 3rd = (20+28)/2 =24 points
4th = (20+28+24)/3 = 24 points
5th = (20+28+24+24)/4 = 24 points

B. 6th = (20+28+24+24+24)/5 = 24 points

C. (X+Y)/2 would be the points scored by the third shooter and all the following shooters.

D. The results would be changed because the third shooter would have added another variable to be averaged. All shooters following the fourth shooter, though would have scored the same points.

Student uses appropriate strategies and produces the correct solutions for the problems.

Student response reflects the ability to reason mathematically and to generalize.

Student response shows an understanding of averaging.

Student can extend the application beyond the given formula and variables.

Student explains how change in one quantity affects change in another quantity.

TABLE 3–23
TESTS ADMINISTERED IN GRADE 4 IN 1994–95 AND 1995–96
(CELL ENTRIES INDICATE TOTAL NUMBER OF UNIQUE ITEMS ADMINISTERED ACROSS FORMS)

Subject Area	Portfolio		Performance Events[2]		Open-Response[3]	
	1994–95	1995–96	1994–95	1995–96	1994–95	1995–96
Mathematics	Grade 5	Grade 5	2	2	29	29
Reading			0	0	29	29
Science			2	2	29	29
Social Studies			2	2	29	29
Writing	●	●	0	0	24	24
Arts & Humanities			3	3	12	24
Practical Living / Vocational Studies			3	3	12	24
Alternate Portfolio	●	●	0	0	0	0
Total Unique Items			12	12	164	188

---

[2] Performance events were administered, but were not included in Accountability Cycle 2 calculations.

[3] Writing open-response questions are different than those for other subject areas. See text for description.

TABLE 3-24
TESTS ADMINISTERED IN GRADE 8 IN 1994-95 AND 1995-96
(CELL ENTRIES INDICATE TOTAL NUMBER OF UNIQUE ITEMS ADMINISTERED ACROSS FORMS)

Subject Area	Portfolio		Performance Events[4]		Open-Response[5]	
	1994-95	1995-96	1994-95	1995-96	1994-95	1995-96
Mathematics	●	●	2	2	30	30
Reading			0	0	30	30
Science			1	2	30	30
Social Studies			2	2	30	30
Writing	●	●	0	0	24	24
Arts & Humanities			3	3	12	24
Practical Living / Vocational Studies			3	3	12	24
Alternate Portfolio	●	●	0	0	0	0
Total Unique Items			11	12	168	192

---

[4] Performance events were administered, but were not included in Accountability Cycle 2 calculations.

[5] Writing open-response questions are different than those for other subject areas. See text for description.

TABLE 3-25
TESTS ADMINISTERED IN GRADE 11 IN 1994-95 AND 1995-96
(CELL ENTRIES INDICATE TOTAL NUMBER OF UNIQUE ITEMS ADMINISTERED ACROSS FORMS)

Subject Area	Portfolio		Performance Events[6]		Open-Response[7]	
	1994-95	1995-96	1994-95	1995-96	1994-95	1995-96
Mathematics	Grade 12	Grade 12	1	2	30	30
Reading			0	0	30	30
Science			2	2	30	30
Social Studies			2	2	30	30
Writing	Grade 12	Grade 12	0	0	24	24
Arts & Humanities			3	3	12	24
Practical Living / Vocational Studies			2	3	12	24
Alternate Portfolio	Grade 12	Grade 12	0	0	0	0
Total Unique Items			10	12	168	192

---

[6] Performance events were administered, but were not included in Accountability Cycle 2 calculations.

[7] Writing open-response questions are different than those for other subject areas. See text for description.

NUMBER OF TEST QUESTIONS PER STUDENT

The technique of matrix sampling was used to increase the coverage of the assessment and to provide data for linking information across years. Matrix sampling divides a large number of test questions among different forms of the test; each student responds to only one form including all of the common items[8] and a fraction of the matrix-sampled items. Table 3–26 presents the number of common and matrix items within each subject taken by each student.

As with previous years, twelve forms of each test were created in 1994–95 and 1995–96. For each subject area, each set of 12 forms covered a total of 24 open-response questions in addition to any common items. For reading, mathematics, science, and social studies, the tests contained five common open-response items in grade 4 and six common open-response items in grades 8 and 11. The common items were publicly released following the administration of the tests, while the matrix-sampled items were kept secure. A subset of matrix-sampled items are reused the following year as common items. After being reused as common items, these items become public with the release of that year's common item pool.

Arts and Humanities and Practical Living/Vocational Studies Items are entirely matrix-sampled, with no common items. In 1994-95 and 1995-96, two items per content area were taken by each student after completion of reading, mathematics, science and social studies. In 1992-93 and 1993-94, a variable number of matrix-sampled Arts and Humanities and Practical Living/Vocational Studies items was taken, typically ranging from zero to two items per student. These 1992-93 and 1993-94 items were integrated into the other content areas, e.g., as a single item covering academic expectations related to both Arts and Humanities and reading presented among the reading items, as there was not sufficient testing time to allow for separate test administrations in Arts and Humanities and Practical Living/Vocational Studies. (Table 3-34 in the KIRIS Accountability Cycle I Technical Manual gives the number of Arts and Humanities and Practical Living/Vocational Studies items integrated into reading, mathematics, science and social studies.)

TABLE 3–26
NUMBERS OF KIRIS OPEN-RESPONSE ITEMS FOR EACH SUBJECT PER STUDENT BY YEAR

Years	Common Items (Reading, Mathematics, Science and Social Studies)		Matrix-Sampled Items (All Grades)	
	Grade 4	Grades 8, and 11[9]	Reading, Math, Science, and Social Studies	Arts & Humanities and Practical Living/Vocational Studies
1992–93, 1993-94	5	5	2	0-2[10]
1994-95, 1995–96	5	6	2	2

---

[8] Common items are those that are repeated in all test forms, and thus are taken by all students.

[9] KIRIS tests were administered at grade 12 in 1992-93 and 1993-94.

[10] In Grade 4 Arts and Humanities for 1993-94, there were three items. Also, in 1993-94, an Arts and Humanities item at grade 12 was inadvertently administered as a common item in reading instead of a matrix item. This item was treated as a matrix-sampled by being counted for accountability purposes only for a single form of the test, form 1.

CHAPTER 4
TEST DEVELOPMENT PROCESS

CONTENT ADVISORY COMMITTEES

The linchpins of the test development process are the content advisory committees. These committees assure that the tests and scoring guides are appropriate for Kentucky students. Advisory committees in each area—reading, mathematics, science, social studies, writing, arts and humanities, and practical living/vocational studies—guide test development for open-response, multiple-choice, performance events, and the writing, mathematics, and alternate portfolios. Members of the committees include Kentucky classroom teachers at each pertinent grade level, resource teachers, administrators, curriculum coordinators, university professors, and Kentucky Education Association representatives. Tables 4–1 and 4–2 summarize the number of committee members in 1994–95 and 1995-96, as well as their backgrounds. (For committee membership information for previous years, see the KIRIS Accountability Cycle I Technical Manual.)

TABLE 4–1
TEST DEVELOPMENT COMMITTEES—MEMBER EXPERIENCE

Year	Content Area	Teachers	Administrators	KDE Staff	College and University
1994-95	Mathematics	18	2	6	3
	Reading	23	4	2	3
	Science	29	2	5	1
	Social Studies	18	1	3	2
	Writing	22	8	13	1
	Arts and Humanities	17	3	2	4
	Practical Living/ Vocational Studies	17	5	2	8
1995-96	Mathematics	21	2	4	3
	Reading	22	5	2	3
	Science	25	2	3	3
	Social Studies	21	3	3	1
	Writing	20	2	11	0
	Arts and Humanities	21	4	2	5
	Practical Living/ Vocational Studies	25	4	3	3

TABLE 4–2
TEST DEVELOPMENT COMMITTEES—MEMBER DEMOGRAPHICS

Year	Content Area	Male	Female	Minority	Non-Minority	Region							
						1	2	3	4	5	6	7	8
1994-95	Mathematics	6	17	3	20	1	5	5	4	4	2	1	1
	Reading	2	28	2	28	6	5	2	2	6	4	2	3
	Science	12	20	3	29	7	4	2	3	9	4	2	1
	Social Studies	8	13	2	19	5	5	4	2	1	1	2	1
	Writing	6	35	3	38	3	4	5	5	9	5	6	4
	Arts and Humanities	6	18	2	22	3	2	5	1	6	5	1	1
	Practical Living/ Vocational Studies	9	21	2	28	3	2	1	6	8	7	3	0
1995-96	Mathematics	10	16	3	23	4	2	3	6	3	4	3	1
	Reading	6	24	3	27	4	6	2	4	5	2	4	3
	Science	11	19	1	29	3	5	1	4	6	6	2	3
	Social Studies	10	15	2	23	4	2	2	4	2	3	3	5
	Writing	4	26	1	29	1	1	3	5	8	3	2	0
	Arts and Humanities	13	17	5	25	3	3	7	4	6	2	3	2
	Practical Living/ Vocational Studies	12	20	4	28	4	4	2	4	6	4	5	3

Note: No KDE staff members are included in reporting member demographics, except for regionally-based Writing Portfolio consultants.

THE TEST DEVELOPMENT SEQUENCE

Test development is an iterative process involving repeated reviews, revisions, and analyses. Throughout this process there is considerable interaction among content advisory committees, consultants from the Kentucky Department of Education, and staff from Advanced Systems in Measurement and Evaluation (or their subcontractors). Table 4–3 presents the major steps in the test development process.

TABLE 4–3
MAJOR STEPS IN THE TEST DEVELOPMENT PROCESS

1.	Analyze score data from previous year's test
2.	Select next year's common items from previous year's matrix items
3.	Select equating items from previous year's matrix items
4.	Develop concepts for new items
5.	Write initial draft of new items
6.	Review and revise initial item drafts
7.	Review and select items for field testing (Content Advisory Committees)
8.	Assemble field test booklets
9.	Review field test booklets
10.	Administer field tests
11.	Analyze field test results
12.	Select new items for upcoming year's test forms
13.	Review selected items (Bias Review Committee)
14.	Review selected items (ASME/KDE staff)
15.	Final item selection (ASME/KDE staff)
16.	Revise selected items and assemble test forms
17.	Review each assembled form for content and sensitivity balance (ASME staff)
18.	Review each assembled form for content and sensitivity balance (KDE staff)
19.	Review final typeset copy

1. **ANALYZE SCORE DATA FROM PREVIOUS YEAR'S TEST.** The last step of one year's test development process is the same as the first step of the next year's process. Item analysis provides information about how well each item performed. This information is used to select the previous years' items that will appear in the following year's test and for making modifications to previous years' items. Table 4–4 describes the primary information produced for open-response items, and Table 4–5 shows a sample of the item analysis output for open-response items. Similar information is produced for multiple-choice items, excepting the biserial correlation between item score and student performance

level (i.e., there are no performance level classifications associated with multiple choice items). Summaries of item analyses are presented in Chapter 6. Members of Content Advisory Committees receive training on the interpretation of score data. The training ensures that each content area committee and individual member understands how information from the analyses is used to select or revise items.

TABLE 4–4
ANALYSES PERFORMED ON OPEN-RESPONSE ITEMS

Analysis	Comments
1. Common item difficulty	Distribution of scores and average item score for the total group of examinees (on 0–4 scale).
2. Form item difficulty	Distribution of scores and average item score for common and matrix items based on group of examinees taking a particular form of the test (on 0–4 scale).
3. Item-total correlation	Correlation between item scores and scores on set of common items. For common items, this correlation is corrected for part-total contamination.
4. $D_{N\text{-}A}$ biserial correlation	Discrimination between Novice and Apprentice. Biserial correlation between item score and student performance level, with performance level dichotomized as either Novice or above Novice.
5. $D_{A\text{-}P}$ biserial correlation	Discrimination between Apprentice and Proficient. Biserial correlation between item score and student performance level, with performance level dichotomized as either Apprentice or below or Proficient or above.
6. $D_{P\text{-}D}$ biserial correlation	Discrimination between Proficient and Distinguished. Biserial correlation between item score and student performance level, with performance level dichotomized as either Proficient or below or Distinguished.
7. DIF-SMD Male-Female	Dorans and Schmitt (1991)* standardized mean difference for differential item functioning, comparing groups of male and female students matched on total common-item scores. A positive DIF-SMD score indicates relative to other items, this item favors males. A negative DIF-SMD indicates relative to other items, this item favors females.
8. DIF-SMD White-Black	Standardized mean difference for differential item functioning, comparing groups of White and Black students matched on total common-item scores. A positive DIF-SMD score indicates relative to other items, this item favors White students. A negative DIF-SMD indicates relative to other items, this item favors Black students.
9. Coefficient Alpha	Internal consistency statistic calculated for common items and for each subform calculated for total set of items within the appropriate content category.

*Dorans, N., & Schmitt, A.P. (1991). *Constructed-response and differential item functioning: A pragmatic approach (ETS research report No. 91–47).* Princeton, NJ: Educational Testing Service.

TABLE 4-5
SAMPLE OPEN-RESPONSE ITEM ANALYSIS OUTPUT

				Frequency distribution				
Form	N	Mean	Std. Dev	0	1	2	3	4
Overall	49704	2.43	.837	1293	2857	23042	17963	4549
1	4174	2.29	.874	118	494	1915	1336	311
2	4139	2.46	.836	111	241	1797	1631	359
3	4152	2.44	.804	99	181	2030	1489	353
4	4165	2.47	.837	106	239	1806	1630	384
5	4158	2.42	.824	118	167	2123	1364	386
6	4175	2.37	.855	117	358	1891	1479	330
7	4136	2.49	.845	113	192	1806	1597	428
8	4160	2.47	.839	107	160	2020	1408	465
9	4142	2.47	.830	117	214	1773	1691	347
10	4120	2.46	.830	87	144	2063	1459	367
11	4088	2.42	.844	96	277	1906	1417	392
12	4095	2.47	.838	104	190	1912	1462	427

TABLE 4–5, continued
SAMPLE OPEN-RESPONSE ITEM ANALYSIS OUTPUT

Form	Corrected Item-Total Corr.	N vs. APD Biserial	NA vs.PD Biserial	NAP vs. D Biserial	Smd -Gender	Smd-Race
Overall	.535	.667	.606	.519	.027	.009
1	.565	.688	.668	.483	*	*
2	.537	.676	.555	.490	*	*
3	.522	.647	.571	.594	*	*
4	.560	.688	.607	.566	*	*
5	.452	.576	.592	.606	*	*
6	.570	.722	.632	.551	*	*
7	.529	.709	.585	.445	*	*
8	.487	.626	.610	.530	*	*
9	.533	.703	.580	.559	*	*
10	.486	.594	.615	.478	*	*
11	.577	.672	.677	.591	*	*
12	.541	.685	.594	.518	*	*

Note: The SMD statistic was not calculated by form for common items.

2. **SELECT NEXT YEAR'S COMMON ITEMS FROM PREVIOUS YEAR'S MATRIX ITEMS.** At their first meeting, the content advisory committees select and make revisions to next year's common items. Decisions are based upon both breadth of coverage of content and academic expectations and statistical quality of the items.

   When committee members review test questions, they sort the items by content categories defined by the content guidelines and academic expectations. Advanced Systems' test developers and the Kentucky advisory committees hold joint responsibility to assure that the subset of questions within a category provide breadth of coverage of the important concepts and skills in that subdomain. With materials organized this way and with approximate quotas for the numbers of items in different categories required in a final test, the developers and committees easily maintain a relatively stable distribution and balance of different items across the categories.

3. **SELECT EQUATING ITEMS FROM PREVIOUS YEAR'S MATRIX ITEMS.** Similarly, at their first meeting the content advisory committees select matrix items that will be re-used to provide the equating link that maintains standards across test forms and years. Decisions are based upon both breadth of coverage of content and academic expectations, statistical quality of the items, and equating needs.

4. **DEVELOP CONCEPTS FOR NEW ITEMS.** The Content Advisory Committees review the existing item pool for areas that need to be replenished. For identified areas, the appropriate Content Advisory Committee generates ideas for

several possible items. The goal is always to develop more items than are needed for each content area. A larger number of items allows the advisory committees more choice and flexibility when selecting new items for the upcoming year's test forms.

5. **WRITE INITIAL DRAFT OF NEW ITEMS.** Content Advisory Committee members receive a resource binder and training to assist their work during the test development process. The binder includes each of the following: development manuals for open response and multiple choice items, and performance events; content information such as Academic Expectations, Content Guidelines, and sections from *Transformations: Kentucky's Curriculum Framework;* and assessment information such as Kentucky's Holistic Scoring Guide, released items from previous KIRIS assessments, and Guidelines for Handling Sensitive Issues in Kentucky's State Assessment Development. Training is designed around the use of the development manuals and other materials found in the binder during the development of assessment items.

   During the first meeting, advisory members are trained and begin generating ideas for new items and/or write initial drafts of new items. After the meeting, ASME staff (typically the test developer assigned to that committee) take the ideas generated by the Content Advisory Committees and develop initial drafts of new items or edit the initial drafts of new items produced by committee members.

6. **REVIEW AND REVISE INITIAL ITEM DRAFTS.** The initial item drafts are reviewed by ASME test development supervisors. Potential improvements are discussed with the item writers.

7. **REVIEW AND SELECTION OF ITEMS FOR FIELD TESTING BY CONTENT ADVISORY COMMITTEES.** At their second meeting, the Content Advisory Committees review items that have been edited or revised by ASME staff, suggest additional revisions, and select items for field testing. As committees make their selection, they consistently examine the item to ensure it reflects Kentucky's Academic Expectations and Content Guidelines.

8. **ASSEMBLE FIELD TEST BOOKLETS.** A test assembler and a test developer select the items that will constitute each field test form.

9. **REVIEW FIELD TEST BOOKLETS.** The assembled field test booklet is reviewed by the appropriate ASME test development supervisor.

10. **ADMINISTER FIELD TESTS.** New questions are field tested, gathering 50 to 100 student responses per question (fewer for performance events), primarily using students from schools in California, Maine, New Hampshire, and Massachusetts.[1]

11. **ANALYZE FIELD TEST RESULTS.** After field-testing, student answers to open-response questions and to performance events are scored by ASME staff using preliminary holistic scoring guides. In addition, ASME staff bring samples of student responses to the next content advisory committee meeting for review. Multiple-choice data are analyzed to produce item statistics used to identify potential problems with items. Questions are revised based on field test results prior to and during committee meetings as the actual student field-test responses are reviewed by committees.

12. **SELECT NEW ITEMS FOR UPCOMING YEAR'S TEST FORMS.** At the third meeting, the Content Advisory Committees from the field test select the remaining test items. Scored data and sample student responses for open

---

[1] Starting with the 1995–96 tests, some item tryout occurred in a separate pretest section of the operational test, rather than in a field test. This allowed larger samples of data based on Kentucky students as the basis of item analysis.

response items and performance events are used by committee members as they select items for the test. Also, the breadth of coverage of content and Academic Expectations provide decision points for selection.

13. **REVIEW OF SELECTED ITEMS BY BIAS REVIEW COMMITTEE.** After final item sets are selected, all items are submitted to the Bias Review Committee. The role of this committee is to ensure that the content of the test is fair and equitable for all students, and that the tests contain no material that could be construed to be racially-, ethnically-, or gender-stereotyped or biased toward any group. Each testing year this committee meets to review all open-response, multiple-choice, and performance event items for every grade level. Committee members use response forms to record their specific concerns about each item, and to provide suggestions for revision.

14. **REVIEW OF SELECTED ITEMS BY ASME/KDE STAFF.** Content Advisory and Bias Review Committees concerns are reviewed, and any appropriate revisions are made.

15. **FINAL ITEM SELECTION BY ASME/KDE STAFF.** Face-to-face meetings with KDE and ASME consultants are conducted. During these meetings, KDE/ASME staff give a final review to the coverage of content and academic expectations and statistical quality of the items. Items and scoring guides are reviewed and any appropriate revisions are made by ASME.

16. **REVISE SELECTED ITEMS AND ASSEMBLE TEST FORMS.** ASME staff build the 12 test forms for each grade level. Performance event sets are finalized.

17. **REVIEW OF EACH ASSEMBLED FORM FOR CONTENT AND SENSITIVITY BALANCE BY ASME STAFF.** While each individual item has been reviewed, the mix of items on each form is now assessed to ensure that content, academic expectations, and politically and socially sensitive issues are in balance within each test form.

18. **REVIEW OF EACH ASSEMBLED FORM FOR CONTENT AND SENSITIVITY BALANCE BY KDE STAFF.** KDE staff review the final draft of each form before signing off on the forms to be typeset. KDE staff must ensure that final revisions agreed upon at the face-to-face meetings have been made to each item.

19. **REVIEW OF FINAL TYPESET COPY FOR POTENTIAL TYPOGRAPHICAL ERRORS.** ASME and KDE staff perform final review on typeset copy before giving approval to print.

SPECIAL STUDIES

Over the course of the second accountability cycle, KDE conducted several special studies as part of on-going test development. A summary of the results of those studies follows.

TEST SENSITIVITY REVIEW BY AD HOC PARENTS AND PUBLIC COMMITTEE. This committee met from June 8–10, 1994 to review all forms of the 1993–94 KIRIS accountability-grade tests for balance and sensitivity to different points of view. The committee consisted of 22 members who were nominated by the Kentucky Council of Churches, Kentucky Education Association, Kentucky Parents Teachers Association, Kentucky School for the Deaf, KIRIS Content Advisory Committees, and citizens who had previously expressed concerns about these issues.

A subset of six individuals expressed strong concerns based on their views that (1) the selection of test questions favored a liberal viewpoint, (2) many questions appear to seek a politically correct response, (3) there is a lack of patriotic and traditional literature used in prompts, (4) religion is often cast in a negative light, (5) multiculturalism is overemphasized,

(6) environmental issues are overemphasized because they are addressed in Science, Vocational Studies/Practical Living, and Social Studies, (7) the feminist viewpoint is given preference over traditionalist views.

Upon discussion, the consensus of the committee was that only one of these concerns—overemphasis of environmental issues—was well founded. Eighteen percent of the open response items (21 out of 116 items) were either about the environment or were set in a context related to the environment. Within the three content areas where these items appeared (Science, Vocational Studies/Practical Living, and Social Studies) the percent was considerably higher. In addition, it appears a disproportionate number of passages for the reading test were somber, bordering on depressing.

Based on these findings, additional balance and sensitivity steps and checks were added to the test development process. The document, Guidelines for Handling Sensitive Issues in Kentucky's State Assessment Development, was developed to provide guidance for those groups who develop test items for the Kentucky assessment.

ANALYSIS OF CONTENT COVERAGE BY NATIONAL EXPERTS. Content coverage of the multiple choice and open-response tests in reading, mathematics, social studies and science for 1993–94 was evaluated following administration of the these tests. For each content area, a review by KDE content area specialists and two national curriculum experts indicated complete content coverage when considered at the school-level, with multiple-choice and open-response items covering every category in the NAEP-based content frameworks. When considered at the student level, multiple-choice items also covered every content category, while open-response items covered all but five of the ninety content framework categories: *Number* in Grade 12 mathematics; *Data Gathering and Reduction* in grade 12 science; and, *Astronomy, Matter, and Geology* and *Natural Resources* in grade 4 science. (For additional information, see *Further Consideration of Issues Related to the Inclusion of Multiple-Choice Items in KIRIS Reported Scores*), available from KDE.

AD HOC COMMITTEE REVIEW OF KIRIS APPROPRIATENESS FOR DEAF AND HEARING IMPAIRED STUDENTS. In October 1994 a committee which included deaf students and parents, teachers, and administrators of deaf and hearing impaired students met to review the KIRIS assessments and make recommendations about the proportion of items that (1) were fully appropriate for deaf students, (2) were appropriate for deaf or hearing impaired students, but for which the difficulty of the item might be affected by factors other than academic achievement in the subject area being assessed, and (3) were not at all appropriate for deaf or severely hearing impaired students. After this review, the committee discussed test administration and appropriate accommodations. Following are some key points from the meeting.

- The majority of the committee felt strongly that although the assessment of reading and writing skills was critically important, having all subject areas so dependent on this form of communication posed a particular challenge and perhaps a barrier to success for deaf and severely hearing impaired students.
- Several items were completely inappropriate for deaf students.
- Several modifications related to vocabulary and syntactic construction were suggested for particular questions. Further discussion led to the suggestion that a video of American Sign Language interpretation might be the most appropriate single accommodation to consider making available.

DEVELOPMENT OF SCORING GUIDES

Scoring guides and training materials for open-response questions that were carried over from Cycle 1 testing remained intact to ensure comparability of scores. New training materials were developed by test development staff and Advanced Systems' scoring staff for those items new to the assessment in 1994–95 and 1995–96.

During initial test development meetings, the advisory committees discussed and drafted scoring guides for open-response questions. These guides were finalized by Advanced Systems' staff and scorers and reviewed by Department of Education content specialists. Scoring guides were then compared to actual papers returned from testing, measuring the expectations for responses against real student work. After this review, scoring guides were revised and refined yet again. Finally, student papers were selected for training packs. Typically, two sets of training packs per question were assembled. One contained clear exemplars of student work at each score point and was used to orient scorers to the range of responses and associated score points for each question. The second pack contained less clear-cut examples of score points and was used in training to gauge readers' internalization of the score points and readiness to begin scoring.

DEVELOPMENT OF MULTIPLE CHOICE ITEMS

Multiple-choice items for the 1995–96 test were included in a combined subjects pre-test section within each form. These items were used as part of a large scale field test and were not included in the accountability index. Multiple choice items for 1993-94 and earlier were likewise not included in the accountability index, but were presented in the same section of the test booklet as open-response items. In each case, multiple choice items were developed following essentially the same procedures described above for open-response items.

CHAPTER 5
SETTING AND VALIDATING PERFORMANCE STANDARDS

Performance standards are set to determine how well students will be expected to perform. Though many of the decisions about performance standards were resolved during the first accountability cycle, three additional studies were conducted during the second cycle. This chapter will review three activities: standard-setting (benchmarking) for the Grade 5 Mathematics portfolio; a study to determine the extent to which Kentuckians agreed with the standards originally set for Mathematics, Reading, Science and Social Studies; and the re-setting of standards for arts and humanities, and for practical living/vocational studies. Extensive documentation of each of these projects is available from the Kentucky Department of Education as a separate, more detailed study.

SETTING STANDARDS FOR MATHEMATICS PORTFOLIOS

BACKGROUND. Because mathematics portfolios were new to the KIRIS assessment in 1992-93, a benchmarking process was needed to define the standards Kentucky expects for students in each grade level. The process was completed for grades 4, 8, and 12 in the spring of 1993, and is described in the Cycle 1 Technical Manual. The Kentucky State Board of Education moved the Mathematics Portfolio from the 4th to the 5th grade for the 1994-1995 school year, thus necessitating the re-setting of standards at the elementary level to identify performance level indicators for grade 5.

Because the standards would be set, and benchmarks identified using actual Kentucky student portfolios, the Grade 5 benchmarking process could not take place until spring 1995, when the student work was available. As a result, teachers could not receive training or score portfolios during the school year. They were given the summer and early fall of 1995 to attend training and score the portfolios, with initial scores due in October, 1995. Now that benchmarking and standard-setting have been completed, teachers will be able to score portfolios within the school year.

The standard setting/benchmark committee for grade 5 Mathematics Portfolios included 18 Mathematics Portfolio Regional Coordinators (4th and 5th Grade), two university faculty members, three mathematics educators in Kentucky, seven Mathematics Portfolio Resource Teachers, two elementary principals, three members of the KDE staff and one Advanced Systems staff member.

PORTFOLIOS USED IN BENCHMARKING. The portfolios used for benchmarking came from Kentucky students. It was important to collect portfolios from across the state and from all performance levels. Therefore, schools were identified based on several criteria including mathematics scores for open response items and portfolios, region, rural/urban location, independent/county affiliation, and performance on the mathematics portion of the open-response test. KDE contacted the principals of 26 schools asking each to identify a teacher to provide 25-30 mathematics portfolios for use in establishing benchmarks and creating a scoring manual. Participating teachers agreed to include in their students' portfolios the two sample tasks disseminated with the "Teacher's Guide Addendum."

BENCHMARKING PROCESS. All committee members were experienced at holistic scoring, having scored 4th and/or 8th grade mathematics portfolios previously. Three days were spent reading many portfolios and rating them according to the language on the Holistic Scoring Guide. Performance level subcommittees recommended 2-3 portfolios at each level that were brought to the entire committee. These portfolios were reviewed and discussed until all members came to a consensus regarding the scoring of the selected portfolios. Those portfolios that best exemplified the language on the Holistic Scoring Guide were selected for inclusion in the teachers' scoring manual. The committee also selected high-end and special consideration portfolios (difficult to score portfolios used for specific training sessions).

KDE and Advanced Systems' staff then wrote rationales explaining the scoring of each portfolio, and completed holistic scoring guides and core concepts sheets, for each portfolio, according to the criteria on the Holistic Scoring Guide. The committee's next step was to reevaluate its scoring decisions and review the rationales. Upon final committee approval, the portfolios and their accompanying rationales were compiled by grade level and published in the KIRIS Mathematics Portfolio Assessment Scoring Training Manuals (Grade 5, Grade 8, and High School).

Each manual contained 8 portfolios: 4 benchmark portfolios (one each Novice, Apprentice, Proficient, and Distinguished), and 4 high-end portfolios (two each Novice and Apprentice). Each type of portfolio served a distinct purpose. Benchmark portfolios provided examples that reflected the "middle" of a particular performance level. High-end portfolios served to assist teachers in making score decisions when scoring difficult portfolios.

STANDARDS VALIDATION FOR MATHEMATICS, READING, SCIENCE & SOCIAL STUDIES OPEN-RESPONSE TESTS

The overall purpose of the standards validation process was to determine whether the Kentucky Instructional Results Information System (KIRIS) performance standards, as originally set in 1992, were appropriate in the judgment of a broad range of Kentucky citizens. In October and November, 1995, more than three hundred Kentuckians, in a study conducted by the Kentucky Department of Education and Advanced Systems, reviewed standards in mathematics, reading, science and social studies at grades 4, 8, and 11, and them found to be appropriate.

The study was undertaken for two reasons. First, KDE had stated that when the standards were first set, the standards would be reviewed after they had been in place for a few years. Second, a report commissioned by the Office of Educational Accountability (OEA) in the Kentucky General Assembly was critical of the original standard setting process, and recommended that the standards be reset.

The National Technical Working Group (NTWG: a group of nationally recognized educational measurement specialists established by KDE to review various aspects of the program and to provide guidance throughout the assessment) effectively argued that the standards should be validated rather than reset; that is, the most logical first step would be to determine whether panels of judges agreed with the standards as originally set and currently in use in Kentucky, rather than approaching the issue as if educators throughout Kentucky had not been using those standards for the past three years. The committee felt that a resetting effort should be undertaken only if the validation process showed the existing standards were substantially out of line with judges' opinions of where they should be set. KDE acted on this recommendation.

VALIDATION PROCESS OVERVIEW. With direction and support from KDE, Advanced Systems designed and implemented the validation study in a two month time frame, involving hundreds of Kentuckians who served as judges. In order to ensure broad representation, KDE used an extensive process — drawing on local districts, professional associations, community and business groups — to identify and invite potential judges. Of the 352 invited Kentuckians, 246 participated as judges.

The first step in the overall standards validation process was to ensure that the general and content-specific definitions were still appropriate. A group comprised of Kentucky Department of Education and Advanced Systems staff reviewed the existing general definitions for the four existing performance levels (Novice, Apprentice, Proficient and Distinguished) and made modifications where appropriate. These definitions were then used as the basis for content-specific definitions.

Advisory groups, consisting of academic subject matter experts, parents, classroom teachers, school administrators, business people and policy makers, developed the content-specific definitions for each content area. One hundred six people, out of the invited 126, participated in this step of the Validation Study. They met in groups to review the general definitions for the four desired performance levels (Novice, Apprentice, Proficient and Distinguished) and then to create subject-specific performance level definitions. These definitions formed the basis for the judgments made by the standards validation panels.

The actual validation process required judges to assign performance levels to actual student work. For their review, judges were provided with sets of student answers to all of the common and some of the matrix-sampled constructed-response questions from the 1994/95 test forms. The judges were asked to match individual sets of student responses to the general and content-specific performance level definitions. Judges were first trained on the definitions for each performance level because they were not to establish their own standards criteria, rather their task was to consistently apply the performance level definitions to the student work they were reviewing.

The validation process, although relying upon individual decisions, included small group and total group discussions. Initially, judges worked individually to review student work and to assign an overall performance level to the work of each student. In the second stage, judges talked with other people at their tables and were given an opportunity to change their judgments as a result of those conversations. Then, performance level classifications from the entire room were tallied and reported to facilitate further discussion. Individual judges again were permitted to make changes in their performance level allocations if they felt the information they now had warranted such a change. Results were generated comparing the judges' final ratings against the performance levels originally assigned to students' work.

In addition to examining the overall appropriateness of the standards, the study was designed to answer two questions: 1) Did opinions vary among the different categories of judges -- teachers, other educators and administrators, parents, policy-makers, and business representatives? and 2) Did opinions vary according to the information provided to judges?

To study the first question, judges were assigned to tables so that all judges at one table belonged to the same category. Data for these groups were recorded separately so that effects due to group membership could be studied. To study the second question, judges were randomly assigned to one of two groups -- the "confirmation" group or the "replication" group. In the confirmation group, judges were told the original classifications; in the replication group, judges were given no such information.

The validation process was first conducted with grade 11 mathematics. Once it was clear that the process had worked well for that content area, the same process was used for the remaining content areas (reading, science and social studies) for grade 11. After the grade 11 standards validation process was completed, the standards for grades 8 and 4 were validated using the same overall procedures, with appropriate modifications for those grades.

IDENTIFICATION OF COMMON AND MATRIX-SAMPLED ITEMS. The process for locating student response sets to be used for the validation study began with the selection of particular test forms from the 1994-95 accountability tests. In grades 8 and 11, students had taken six common and two matrix-sampled questions; in grade 4, five common and two matrix-sampled questions. The design team decided to select student work across four forms for each grade, thereby providing judges with samples of student work across 13 (for grade 4) or 14 (for grades 8 and 11) test questions, per content area. It was felt that these numbers would give judges an opportunity to see a broad range of student work, while at the same time minimizing the number of questions and training guides they would need to review. Because judges would see only a subsample (30 for grades 8 and 11 or 29 for grade 4) of the questions in the test, the four forms chosen for each academic area were selected to reflect as a set, broad content coverage.

For each content area within each accountability grade, a set of training materials was compiled that contained each identified item and the associated descriptor of a response receiving a score of 4 points (the highest possible). The inclusion of the "4" descriptor enabled judges to get a sense of the criteria scorers used for a complete response to each item. The other score points were not included because the purpose of the validation work was not to have judges score each item but to align student work with general and content-specific performance level definitions.

IDENTIFICATION OF STUDENT RESPONSE SETS. The next step in the validation process was the compilation of lists containing logit values of student responses to item sets for all identified forms in descending order according to logit values. Logits are item response theory-based ability estimates ranging from about -3.0 to +3.0. Logits correlate very highly with raw score totals. Once the lists were compiled, random samples of 50 students per quarter logit in each content area in grade 11, and 25 students per quarter logit in grades 4 and 8, were identified.

For each academic area and grade level, KIRIS has determined values, in terms of the logit score for each student, that establish the cut points between adjacent performance levels. From these cut points and the highest and lowest logit scores, the midpoints were identified for the four performance levels. Using the random sampled student response set lists and the cut points and midpoint data for each academic area, samples of five students were selected to represent each of seven categories: Novice, Novice/Apprentice, Apprentice, Apprentice/Proficient, Proficient, Proficient/Distinguished and Distinguished. Student responses were eligible to be selected if their logit score was within 0.1 of the appropriate cut point or midpoint and there were no blank, off-topic, or illegible responses to any of the questions. (An equating error that was corrected after papers were selected later revealed that students' responses varied slightly in their distance from the cutpoint or midpoint. Details of the correction, which had essentially no impact on the results of this study, can be found in the complete study, *KIRIS Standards Validation Report*.)

RESULTS. The results of the study involved a comparison of the judges' ratings of students' work with the way the work had been classified originally as a result of the initial standard setting conducted in 1991-92, and the equating that had been done to maintain those standards in the years since. Tables 5-1 through 5-.4 provide the overall results of the validation study. Tables 5-1 and 5-2 provide the results for the confirmation group, and tables 5-3 and 5-4 give similar results for the replication group. For each group, the first table shows the median judgment. The second table displays the percentage of cases in which judges thought the students were appropriately categorized by the original standards that had been set ("Standards should be same"), were categorized too low ("Standards are too low"; e.g., work categorized as Apprentice was work they felt was actually Proficient), or were categorized too high. A case is defined by a packet of student responses for each content area within a grade.

TABLE 5-1
MEDIAN JUDGMENTS FOR CONFIRMATION GROUP

Current Standards	Grade 4				Grade 8				Grade 11			
	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies
N	N	N	N	N	N	N	N	N	N	N	N	N
A	A	A	A	A	A	A	A	A	A	A	A	A
P	P	P	P	P	P	P	P	P	P	P	P	P
D	D	D	D	D	D	D	D	D	D	D	D	D

TABLE 5-2
PERCENTAGE OF CASES JUDGED IN CONFIRMATION GROUP TO AGREE WITH STANDARDS

Standards should be:	Grade 4				Grade 8				Grade 11			
	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies
Lower	1	11	2	3	1	0	7	7	13	2	5	3
Same	86	66	70	79	89	79	78	59	61	82	77	82
Higher	14	23	28	18	10	21	15	34	26	17	17	15

TABLE 5-3
MEDIAN JUDGMENTS FOR REPLICATION GROUP

Current Standard	Grade 4				Grade 8				Grade 11			
	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies
N	N	N	N	N	N	N	N	N	N	N	N	N
A	A	A	A	A	A	A	A	A	A	A	N	A
P	P	P	P	P	P	P	P	P	P	P	A	A
D	D	P	P	P	D	D	P	D	D	D	P	D

TABLE 5-4
PERCENTAGE OF CASES JUDGED IN REPLICATION GROUP TO AGREE WITH STANDARDS

Standards should be:	Grade 4				Grade 8				Grade [illegible]1			
	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies	Reading	Math	Science	Social Studies
Lower	0	4	6	4	5	12	2	9	12	5	2	3
Same	73	70	69	68	80	74	67	60	68	80	50	67
Higher	27	26	25	28	15	14	31	31	20	15	48	30

As can be seen from the tables, judges generally agreed with the standards as they were originally set. For the confirmation group, their median judgment in every case was that middle-range Novice work was indeed Novice, middle-range Apprentice work was indeed Apprentice, and so on. When there was disagreement, the tendency was to believe that the quality of the work shown was not as high as it had been categorized. For every grade level in every subject, the majority of judges felt that student work was appropriately categorized. Of those who did not side with the majority, more felt that the standards should be raised than felt they should be lowered.

The replication group had the same general tendencies, although a higher percentage in this group thought the standards should be higher than did the confirmation group. In contrast to the confirmation group, where at every grade in every subject, the median judgment was that the standards were correct as originally assigned, there were some cases for the replication group in which they thought the work they were reviewing was not as good as it had been originally categorized. Most of these disagreements came when they were looking at work that had been originally categorized as Distinguished, but which they felt was only Proficient work. In addition, there was a feeling among this group that the work they were looking at for grade 11 science was categorized too highly, but this result may have been the consequence of a few influential judges, as shall be shown later in this section. Consistent with the above finding, it can be seen from Table 5-4 that a smaller percentage of judges in the replication group agreed that the standards should remain as they were than was true for the confirmation group; and again, for every subject in every grade, when the replication judges disagreed, they felt that the standards should be raised, not lowered.

RESULTS FOR THE PARENT, POLICY MAKER, BUSINESS PEOPLE ,TEACHER AND OTHER EDUCATOR GROUPS.

As was noted in an earlier section, three different categories of people participated as judges at grades 4 and 8, with two additional categories added for grade 11. This was done because there was concern that different groups might have different familiarity with the current standards, and also because these various groups might have different beliefs about how high the standards should be. One of the criticisms of the original standard-setting process was that the judges were too limited in their backgrounds, and inviting a variety of judges for this study was a way to respond to this concern.

To summarize the results of the various groups, both the original categorizations and the ratings made by the judges were placed on a scale from 1 to 7, where 1 represented mid-range Novice work, 2 represented Novice/Apprentice borderline work, 3 represented mid-range Apprentice work, and so on, with 7 representing mid-range Distinguished work. From this recoding process, judges' ratings can be reported as a discrepancy between the original categorization and the rating. In the system used in this report, a positive value for a discrepancy score means that the judge's rating was higher than the original categorization (or, put another way, that the judge thought the standard was too high, and the judge's opinion would be that the work should be placed in a higher category). Similarly, a negative discrepancy score would mean that the judge felt the standard was too low, and the judge would not have categorized the students' work as highly.

Table 5-5 provides the mean discrepancy scores for the various groups.

TABLE 5-5
JUDGES' MEAN DISCREPANCY SCORES, REPORTED BY GROUP

Grade	Group	Reading	Math	Science	Social Studies
4	Other Educators	-0.39	-0.50	-0.18	-0.59
	Teachers	-0.32	-0.69	-0.46	-0.40
	Parents	-0.45	-0.48	-0.65	-0.38
8	Other Educators	-0.23	-0.14	-0.56	-0.42
	Teachers	-0.18	-0.44	-0.55	-0.61
	Parents	-0.31	-0.10	-0.12	-0.45
11	Business	-0.21	-0.21	-0.24	-0.10
	Other Educators	-0.17	-0.62	-0.76	-0.17
	Teachers	-0.21	-0.26	-0.65	-0.34
	Parents	-0.16	-0.30	-1.30	-0.19
	Policy Makers	-1.10	-0.06	-0.08	-0.76

Only one pattern seems to emerge from Table 5-5. In every cell, the value of the mean discrepancy is negative, meaning that judges thought that student work, on average across all the samples they saw, was placed in too high a category, or that they thought standards were too low. Given that the difference between two adjacent standards is 2 points, these negative values generally are small, but the consistency of the finding is clear.

The groups that had the greatest and least negative values varied across content areas and grade levels. Perhaps in part because there was considerable overlap in the roles of the judges (judges who were invited because they were teachers often were parents as well, for example), the group to which judges belonged appeared to have little influence on their satisfaction with the existing standards. By contrast, there were consistent differences between the judges who were told how the papers had been categorized (the "confirmation" group), and those who were not given this information (the "replication" group). The difference between the groups in these two conditions can be seen from Table 5-6.

TABLE 5-6
JUDGES' MEAN DISCREPANCY SCORES, REPORTED BY CONDITION

Grade	Condition	Reading	Math	Science	Social Studies
4	Confirmation	-0.18	-0.38	-0.39	-0.23
	Replication	-0.65	-0.68	-0.58	-0.83
8	Confirmation	-0.16	-0.41	-0.11	-0.40
	Replication	-0.30	-0.15	-0.70	-0.63
11	Confirmation	-0.49	-0.31	-0.19	-0.17
	Replication	-0.31	-0.30	-1.20	-0.52

In 9 of the 12 cases, the replication group judged the student work as being further below the existing standards than the confirmation group. Note once again that judges do not have strong disagreement with the standards, but if anything, they would make the standards higher. These data show that telling the judges how the papers were originally categorized made them more accepting of that categorization.

Interestingly, although the mean discrepancy scores varied somewhat from group to group and content area to content area, there rarely was disagreement within a table. Given the process used to obtain the ratings, this is a curious result. There were twelve completely separate sessions, one for each combination of grade and subject area. No judge was involved in more than one session, and no information from any one session was distributed to any other session. Within a session, judges were assigned to tables, with all the judges at one table belonging to the same group and condition. Judges were asked to evaluate each of the packets of student work without input from any of the other judges. After they had done that, however, they were encouraged to share their judgments with the other people at their table, and then there was conversation across tables. Judges recorded their initial evaluations, but then were free to change those evaluations if conversations with other judges were persuasive. There was space on the form for three sets of judgments, to allow for the possibility that judges would change their evaluations after each activity.

The first immediate observation from the data was that judges rarely changed their ratings. In the vast majority of cases, only one rating was recorded on the form. It is possible that some judges did not record their evaluations until the end of the session, and thus changed their minds without properly recording that they did so, but the project staff that observed the sessions did not believe this happened often at all. Thus, one might expect the variance of judges within tables to be not much less than the total variance of judges. This was not the case, however. Judgments within tables were highly uniform, suggesting that more communication was occurring before initial evaluations were recorded than the observers reported.

For example, the greatest negative mean discrepancy score for any table at grade 11 came from the table of "replication parents" in science. Their mean discrepancy score was -1.97. Theirs was one of only three tables to have a mean discrepancy score less than -1.00. Yet every person at that table had a mean discrepancy of less than -1.00. Three of the four people at that table had the most negative mean discrepancy scores of anyone in the entire room. A possible *explanation for this result would be that science* standards were set especially low. That does not appear to be a reasonable explanation, however, since student performance (relative to the standards) is lowest in science at every grade level in Kentucky. Among the tables in the full report is a set that provides the standard deviation of mean deviation scores by table. Those tables show just how extreme this result was. As a result, it appears that it would be appropriate to consider the number of tables, not the number of raters, as the unit of analysis in further studies.

RESULTS FOR THE VARIOUS SCORE POINTS. All of the results provided to this point are averages across all the papers provided to the committees. An additional issue that was investigated was whether the general finding presented above (raters generally agreed with the standards, but when they disagreed, they thought the standards should be raised) is true across all standards, or true for some but not for others. In order to answer this question, means of scorer ratings were computed for each of the groups of papers presented. Table 5-7 illustrates these computations for the data across all scorers for grade 11 social studies.

TABLE 5-7
MEAN RATINGS FOR GRADE 11 SOCIAL STUDIES
REPORTED BY KIRIS CLASSIFICATION

"Key" (Classification of Student's Work by KIRIS)	Mean Rating of Students' Work by Raters
1—Mid-range Novice	1.04
2—Novice/Apprentice borderline	1.78
3—Mid-range Apprentice	2.64
4—Apprentice/Proficient borderline	3.62
5—Mid-range Proficient	4.34
6—Proficient/Distinguished borderline	5.51
7—Mid-range Distinguished	6.47

Had the judges perfectly agreed with the classifications already set by KIRIS, their mean rating would be 1.00 for the first row, 2.00 for the second row, and so on. Note that their mean for the first row was very close to 1.00 (1.04 mean); this indicates that almost every judge who was shown mid-range Novice work rated it as Novice, whether or not the judge was informed that the work had been categorized as Novice by KIRIS. For the categories between 2 and 6, a higher rating by one judge could balance off a lower rating by another.

One can see from Table 5-7 that, for grade 11 social studies, there was consistent agreement about midrange work and borderline work. Mid-range Novice work was given mean ratings that were almost exactly on target (1.00), while mid-range Apprentice, mid-range Proficient and mid-range Distinguished work were given means that were lower by .36, .66, and .53, respectively. The borderline Novice/Apprentice work was lowered by .22 by the judges, borderline Apprentice/Proficient work was rated down .38, and borderline Proficient/Distinguished work was lowered by .49. Note that the means would have remained on target if judges agreed with the standards (because there would have been as many judgments to lower a score as to raise it). Thus, a reasonable generalization from these data is that judges had a tendency to feel that mid-range and borderline work belonged in the lower category, not the higher one. This result for grade 11 social studies was typical. Similar results were reached for every grade level, subject area, condition, and affiliation of judge; these results are included in the full report which is available from KDE.

SUMMARY OF FINDINGS. The following is a summary of results from this validation process.

- Judges generally agreed that the standards being used in KIRIS are appropriate; but when they disagreed a substantial minority indicated that the standards should be raised.
- There was not a consistent difference when judges were grouped by affiliation (teachers, other educators, parents, business people, policy makers).
- There was a consistent difference in ratings when judges were told how the student work was originally classified by KIRIS. Those who were told tended to agree with the original classification more than those who were not told.

STANDARDS SETTING FOR ARTS & HUMANITIES, PRACTICAL LIVING/VOCATIONAL STUDIES OPEN-RESPONSE QUESTIONS

This report presents the results of a study conducted in October and November of 1995 to set the standards used in KIRIS for the accountability grades in arts & humanities, practical living/vocational studies. Standards were originally set in 1993 for these content areas but, because of the limited number of matrix-sampled items taken by a student and the integrated nature of these items in the first two years of testing, the National Technical Working Group advised the Kentucky Department of Education that the standards needed to be reset.

STANDARDS SETTING PROCESS OVERVIEW. With direction from KDE, Advanced Systems designed and implemented the standards setting study in a two month time frame, involving more than a hundred people across the Commonwealth as judges. In order to ensure broad representation, KDE used an extensive process -- drawing on local districts, professional associations, community and business groups -- to identify and invite potential judges. Of the 175 invited Kentuckians, 130 participated as judges.

The first step in the overall standards setting process was to ensure that the general and content-specific definitions were still appropriate. A group comprised of Kentucky Department of Education and Advanced Systems staff reviewed the existing general definitions for the four existing performance levels (Novice, Apprentice, Proficient, Distinguished) and made modifications where appropriate. These definitions were then used as the basis for content-specific definitions.

Advisory groups, consisting of academic subject matter experts, parents, classroom teachers, school administrators, business people and policy makers, developed the content-specific definitions for each content area. (This step was conducted concurrent with the Standards Validation Study, which was completed within the same time frame and which is described in the report of that study.) These definitions formed the basis for the judgments made by the standards setting panels. Copies of these definitions may be found in the full report.

The standard setting process required judges to individually review student work and assign an overall performance level. Judges were assigned to tables so that all judges at one table belonged to the same group (teacher, educator, parent, business leader and policy maker). The collection of papers were given to the judges in randomized order so as to minimize information about where a given set of examples of student work might be placed with regard to performance level classification. In the second stage of the study, judges talked with other people at their table and were given an opportunity to change their judgments as a result of those conversations. In the last phase of this process, performance level classifications from the entire room were tallied and reported in order to facilitate further discussion. Individual judges again were permitted to make changes in their performance level allocations if they felt that the information they now had warranted such a change. Following this third step, judges turned in their forms indicating their final decisions about the level of work of each student. Results were generated by establishing cut points determined from judges' final ratings for students' work on KIRIS.

The standards setting process for Grade 11 was completed first, after which the standards for grades 8 and 4 were set using the same overall procedures, with appropriate modifications for those grades.

IDENTIFICATION OF MATRIX-SAMPLED ITEMS. The process for locating student response sets to be used for the standards setting study began with identification of all unique arts and humanities and practical living/vocational studies matrix sampled items used in the 1992/93, 1993/94 and 1994/95 test forms for grades 4, 8 and 11. Open-response tests were administered to grade 12 students in 1992/93 and 1993/93. Because of the limited number of items in these areas all items were used to ensure adequate coverage.

For each content area within each accountability grade, a set of materials was compiled that contained each test item and the associated descriptor of a response receiving a score of 4 points (the highest possible). The inclusion of the "4" descriptor enabled judges to get a sense of the criteria scorers used for a complete response to each item. The other score points were not included since the purpose of the standards setting work was not to have judges score each item but to align student work with general and content-specific performance level definitions. The full report includes a sample of student work for each of the four performance levels (Novice, Apprentice, Proficient and Distinguished) for each accountability grade.

IDENTIFICATION OF STUDENT RESPONSES. The next step in the standard setting process was the compilation of lists containing 20 item score points falling approximately equidistant from each other on the log-odds ability scale for the tests equated across testing years. For each item score point selected, a packet of 10 samples of student work reflecting that score for that item was compiled. An initial random sample of 20 student responses for each score point was identified.

RESULTS. The data from this study were treated quite differently from those collected in the validation study done for reading, mathematics, science and social studies. In that study, standards already were established and in use, and the primary question asked was whether the standards in use were appropriate. Therefore, the basic analysis of data in that study involved the comparison of judges' ratings against the way the student work had been categorized by the existing standards.

Such standards had not been widely promulgated for the two content areas included in this study, in part because the testing program added these content areas a year after the testing program had begun with the other content areas. As a result, the study was approached as a standards setting exercise, rather than standards validation as had been the case for reading, mathematics, science and social studies.

Rather than being shown student work that was categorized in a particular way, judges in this study were shown 20 samples of student work that ranged from low to high in fairly constant scaled score intervals. This collection of papers was given to the judges in randomized order so as to minimize information about where a given set of examples of student work might be placed with regard to performance level classification. The only information provided was the score point assigned to a particular set of papers. By contrast, the scales used in analyzing the results were the log-odds scale item difficulty, and the 1 to 7 scale in which Novice, Apprentice, Proficient and Distinguished were assigned values, 1, 3, 5 and 7 and the intermediate values were assigned to cutpoints between adjacent performance levels.

Many of the 20 samples were taken from different questions; some of the different points along the scaled scores were work on the same question, but assigned a different score (SAS, the statistical analysis program used to create the scaled score points, computes a different "item difficulty" value for each raw score point; higher raw scores have higher item difficulties than do lower raw scores). The purpose of the data analysis was to determine the scaled score point where judges' ratings shifted from being predominately in one category to being in another.

The first issue to be studied was whether there was any significant relationship between judges' ratings and the scaled scores assigned to the student work. If the correlations were too low, further analysis would be fruitless. That is, it would not be possible to find a meaningful point along the item difficulty scale where judges' ratings changed from one category to another if there was no relationship between the two variables. The higher the correlation, the more appropriate this task would be.

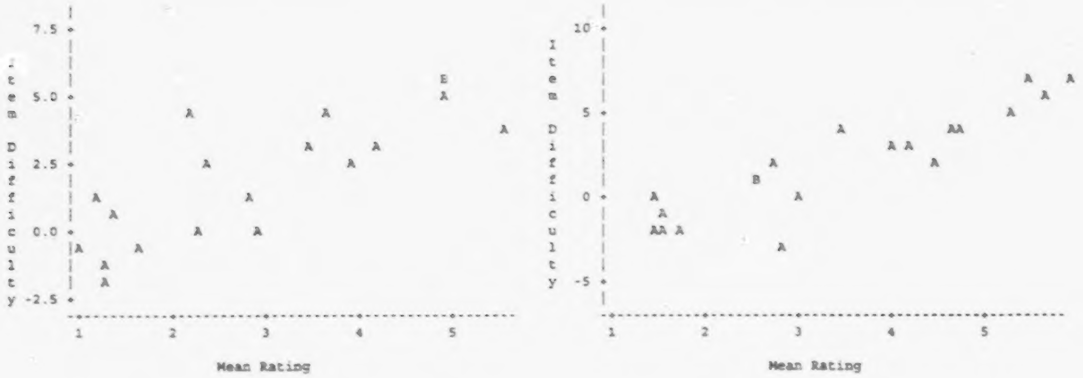
To explore this issue, the mean of judges' ratings was computed for each student's work sample. Judges' ratings of Novice to Distinguished were converted to a 7-point scale, just as had been done in the validation study for reading, mathematics, science and social studies. Judgments of "Novice" were converted to a score of "1", "Novice/Apprentice borderline" was converted to a "2", "Apprentice" to a "3", and so on. After these conversions had been completed, a mean rating was computed for each piece of student work. These mean ratings were compared to the item difficulties (log-odds ratios). Scatterplots showing the relationships between the mean ratings and the item difficulties are provided in Figure 5-1. As can be seen from the scatterplots, the relationships between the two sets of statistics were quite high and linear. As given in Table 5-8, the correlations ranged from a low of .81 to a high of .90.

An interesting side issue that the scatterplots brought to light is best exemplified by the scatterplot for Practical Living/Vocational Studies for grade 8. While the correlation between item difficulty and mean ratings is quite high (.83), few of the data points fall along the regression line; most are either well above or well below the line. Looking at the scatterplot, it would be easy to believe that there actually are two relationships—one representing the points that would be above the regression line, and a second for the points below. However, looking at the questions yielded no reasonable explanation for this phenomenon. In fact, as often as not, the points on either side of the regression line came from the same question, but were just different raw score points from those questions.

81

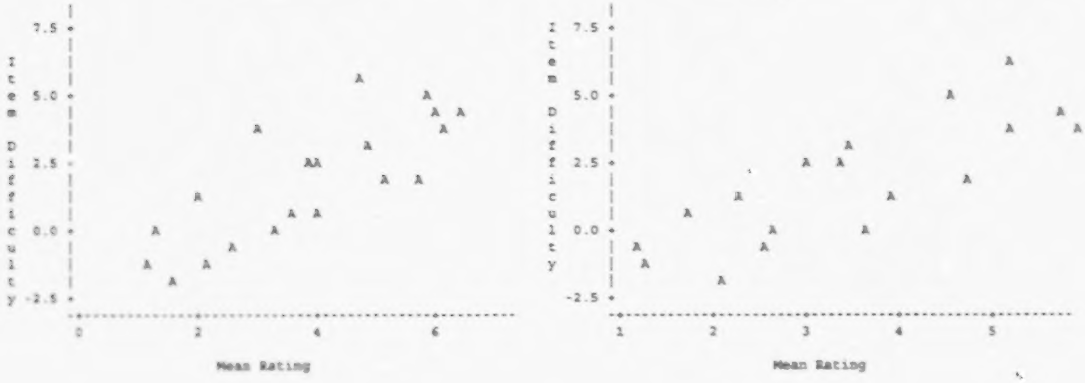
FIGURE 5-1
SCATTERPLOTS OF SCALED DIFFICULTY *VS.* MEAN RATING

Grade 4


Arts and Humanities

Practical Living/Vocational Studies

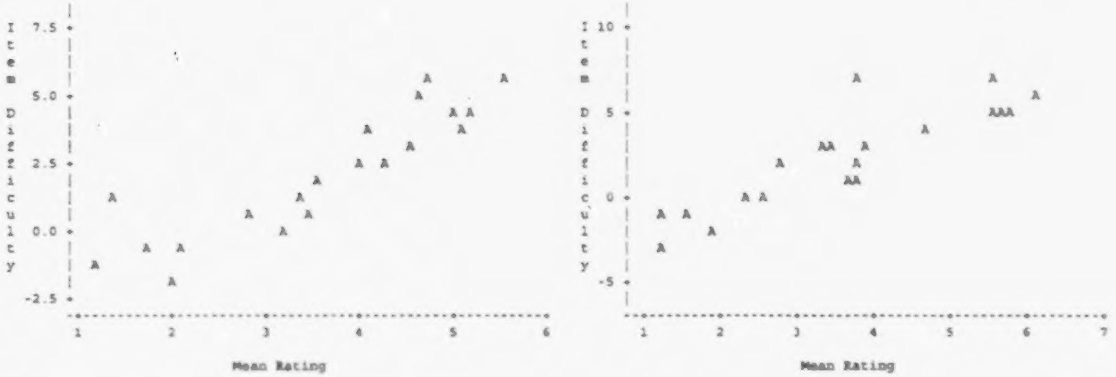
Grade 8


Arts and Humanities

Practical Living/Vocational Studies

FIGURE 5.1, CONTINUED

Grade 11



Arts and Humanities

Practical Living/Vocational Studies

TABLE 5-8
STATISTICS RELATED TO THE SCATTERPLOTS IN FIGURE 5-1

			Regression Line	
Scatterplot		Correlation	Slope	Intercept
Grade 4	Arts and Humanities	.84	1.396	-2.050
	Practical Living/ Vocational Studies	.90	1.815	-4.468
Grade 8	Arts and Humanities	.81	1.127	-2.521
	Practical Living/ Vocational Studies	.83	1.253	-2.544
Grade 11	Arts and Humanities	.90	1.579	-3.570
	Practical Living/ Vocational Studies	.89	1.674	-3.806

Given the regression lines provided in Table 5-8, it became a straightforward process to determine the difficulty levels that are at the cut points; the Novice/Apprentice cut point simply is a mean judges' rating of 2, the Apprentice/Proficient cut point a mean rating of 4, and the Proficient/Distinguished cut point a mean rating of 6, all by definition. For example, to find the appropriate cut point for Novice/Apprentice for grade 4 arts and humanities, one starts with the equation:

Item Difficulty = 1.396 * Mean Rating - 2.050

By inserting the value of "2" for "Mean Rating," one obtains the result that a grade 4 student needs to achieve a scaled score of at least 0.742 to be judged as Apprentice rather than Novice in arts and humanities. The comparable figures for cut-points in all grades and subject areas are provided in Table 5-9.

TABLE 5-9
SCALED SCORES THAT DEFINE THE CUT POINTS

Grade and Content Area		Scaled Score Needed to be Judged as at Least:		
		Apprentice	Proficient	Distinguished
Grade 4	Arts and Humanities	0.742	3.534	6.326
	Practical Living/ Vocational Studies	-0.839	2.790	6.419
Grade 8	Arts and Humanities	-0.267	1.987	4.241
	Practical Living/ Vocational Studies	-0.038	2.468	4.974
Grade 11	Arts and Humanities	-0.411	2.746	5.904
	Practical Living/ Vocational Studies	-0.458	2.890	6.238

This section has outlined the process that was used to set the standards used in the Kentucky Instructional Results Information System for the accountability grades in arts and humanities, and practical living/vocational studies. Readers wishing more detailed information are referred to the full report, available from KDE.

CHAPTER 6
ITEM ANALYSIS

OVERVIEW

To provide evidence of the technical quality of the KIRIS assessment, a series of item level analyses were performed for each KIRIS accountability grade. The results presented in this chapter are from the actual statewide administration of questions for Accountability Cycle 2, the years 1992-93 through 1995-96. The following summarizes the six types of analyses presented for each accountability grade:

- Distribution of item scores for open-response items,
- Distribution of corrected item-total correlations for open-response items,
- Distribution of item-theta correlations for open-response items,
- Distribution of N vs. A, P, D biserial correlations for open-response items,
- Distribution of N, A vs. P, D biserial correlations for open-response items,
- Distribution of N, A, P vs. D biserial correlations for open-response items.

To provide a comprehensive overview for each of the above analyses, distributions of results for each analysis are presented. While this leads to the presentation of an enormous amount of statistical results, it enables the reader to more thoroughly evaluate the technical qualities of KIRIS items.

GRADE 4 RESULTS. Tables 6-1 through 6-6 present the results for the item level analyses for grade 4. Table 6-1 presents the distribution of item scores for open-response items for mathematics, reading, science, social studies, arts and humanities and practical living for each year of Accountability Cycle 2. Note that KIRIS open-response items are scored on a 0-4 scale, with 2.00 representing the middle of the difficulty range. Inspection of Table 6-1 shows that for each year, most of the item difficulties fall between 1.21 and 2.00. In general, the mathematics and science tests across the four years were more difficult than the other four tests. This latter result is reflected in the greater number of mathematics and science item difficulties falling below 1.21.

Results with respect to the corrected item-total correlations presented in Table 6-2 are generally very favorable. Most of the corrected item-total correlations for each test, for each year, generally fall within the range of .40 to .59. It should be noted that the calculation of the corrected item-total correlations for the common items in mathematics, reading, science, social studies was slightly different than for the matrix items. The strategy for obtaining values for the common items was to calculate the corrected item-total correlations for both the common and matrix items by form, and then take the average corrected item-total correlation for each common item across all forms. Using this strategy, each item (both common and matrix) contributed a single value to the distribution of item-total correlations for each content area and year. (Of course, the item score was removed from the corresponding total before the corrected item-total correlation was calculated for an item.)

85

Because all arts and humanities and practical living items were matrix items, the correlations presented for these items actually represents the correlation of one matrix item with another (there being only two matrix items per form for each of these tests). As such, the correlations for arts and humanities and practical living are slightly lower than for the other four test, values generally ranging between .30 to .49. Unlike the results presented in Table 6-2, the item-theta correlations presented in Table 6-3 take advantage of the Item Response Theory (IRT) scale constructed for a content area for a particular year. Because theta is the IRT ability estimate computed for each student based on his or her pattern of responses to questions in a given content area, the correlations presented in Table 6-3 are not corrected. Inspection of Table 6-3 reveals that the correlations are generally higher than the corrected raw score item-total correlations, most item-theta correlations falling within the range of .60 to .79.

To determine the degree of item *discrimination* relative to different student performance levels (i.e., Novice, Apprentice, Proficient, and Distinguished; or N, A, P, D), biserial correlations were obtained between item scores (i.e., 0-4 scale) and dichotomized performance levels so as to assess the relationship between item scores and the assignment of students to a specific performance category. The biserial correlation can be interpreted as an estimate of the product-moment correlation that would have been obtained had the dichotomized performance levels been a normally distributed continuous measure. With respect to the biserial correlations, inspection of Tables 6-4 through 6-6 shows that most items differentiated between Novice and non-Novice students reasonably well and between Apprentice or below and Proficient or above students reasonably well; the biserial correlations in both cases generally fall between .60 and .79. While the biserial correlations for the distinction between Distinguished and non-Distinguished students (Table 6-6) were generally not as high as for the previous two comparisons, they to suggest that most items generally differentiated between Distinguished and non-Distinguished reasonably well; the biserial correlations generally fall between .50 and .79. Note that for some forms of the test, no students achieved a performance level of Distinguished. For these forms, no biserial correlation could be calculated.

GRADES 8 AND 11/12 RESULTS. Tables 6-7 through 6-18 present the results for the item level analyses for grades 8 and 11/12. Because the results for these KIRIS accountability grades were generally similar to those of grade 4, details of the results as presented above are not outlined here. However, several observations with regard to item scores are noteworthy. First, inspection of Tables 6-7 and 6-13 show the grade 8 and 11/12 items to be generally more difficult than for grade 4, with more items across all content areas and years falling below 1.21. Similar to grade 4, the mathematics and science tests across the four years were generally more difficult than the other four tests. However, unlike grade 4, mathematics for grades 8 and 11/12 appeared to have more difficult items than the science test, with more item means falling below .81 (on a scale from 0-4), especially for grades 11/12.

TABLE 6-1
DISTRIBUTION OF ITEM SCORES FOR OPEN-RESPONSE ITEMS, GRADE 4

Subject Area	Year	Item Scores										N Items
		.00-.40	.41-.80	.81-1.20	1.21-1.60	1.61-2.00	2.01-2.40	2.41-2.80	2.81-3.20	3.21-3.60	3.61-4.00	
Mathematics	1992-93	0	0	2	12	14	1	0	0	0	0	29
	1993-94	0	0	5	12	10	2	0	0	0	0	29
	1994-95	0	0	3	10	12	4	0	0	0	0	29
	1995-96	0	0	3	15	8	3	0	0	0	0	29
Reading	1992-93	0	0	0	10	17	2	0	0	0	0	29
	1993-94	0	1	2	10	14	1	1	0	0	0	29
	1994-95	0	0	1	3	13	10	2	0	0	0	29
	1995-96	0	0	0	2	16	10	1	0	0	0	29
Science	1992-93	0	0	0	5	18	5	1	0	0	0	29
	1993-94	0	2	7	4	13	2	1	0	0	0	29
	1994-95	0	0	1	12	11	5	0	0	0	0	29
	1995-96	0	0	5	10	12	2	0	0	0	0	29
Social Studies	1992-93	0	0	0	5	16	7	1	0	0	0	29
	1993-94	0	0	0	10	11	8	0	0	0	0	29
	1994-95	0	0	0	6	17	6	0	0	0	0	29
	1995-96	0	0	0	10	18	1	0	0	0	0	29
Arts and Humanities	1994-95	0	0	2	5	3	2	0	0	0	0	12
	1995-96	0	0	0	13	7	3	1	0	0	0	24
Practical Living	1994-95	0	0	2	3	7	0	0	0	0	0	12
	1995-96	0	0	1	7	15	1	0	0	0	0	24

TABLE 6-2
DISTRIBUTION OF CORRECTED ITEM-TOTAL CORRELATIONS FOR OPEN-RESPONSE ITEMS, GRADE 4

Subject Area	Year	Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	0	11	18	0	0	0	0	29
	1993-94	0	0	0	2	5	17	5	0	0	0	29
	1994-95	0	0	0	0	4	14	11	0	0	0	29
	1995-96	0	0	0	0	4	19	6	0	0	0	29
Reading	1992-93	0	0	0	0	0	20	9	0	0	0	29
	1993-94	0	0	0	0	3	20	6	0	0	0	29
	1994-95	0	0	0	0	2	24	3	0	0	0	29
	1995-96	0	0	0	0	1	24	4	0	0	0	29
Science	1992-93	0	0	0	1	13	15	0	0	0	0	29
	1993-94	0	0	0	3	10	16	0	0	0	0	29
	1994-95	0	0	0	1	15	13	0	0	0	0	29
	1995-96	0	0	0	3	21	5	0	0	0	0	29
Social Studies	1992-93	0	0	0	0	10	19	0	0	0	0	29
	1993-94	0	0	0	0	9	19	1	0	0	0	29
	1994-95	0	0	0	0	4	18	7	0	0	0	29
	1995-96	0	0	0	0	10	16	3	0	0	0	29
Arts and Humanities*	1994-95	0	0	0	6	6	0	0	0	0	0	12
	1995-96	0	0	0	5	7	0	0	0	0	0	24
Practical Living	1994-95	0	0	0	5	7	0	0	0	0	0	12
	1995-96	0	0	1	2	9	0	0	0	0	0	24

* Note: only one correlation per form is presented for arts & humanities and practical living|vocational studies, as the corrected item-total correlation is between two items appearing on the same form.

TABLE 6-3
DISTRIBUTION OF ITEM-THETA CORRELATIONS FOR OPEN-RESPONSE ITEMS, GRADE 4

Subject Area	Year	Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	0	0	5	13	11	0	0	29
	1993-94	0	0	0	0	2	3	10	14	0	0	29
	1994-95	0	0	0	0	1	3	10	15	0	0	29
	1995-96	0	0	0	0	0	4	14	11	0	0	29
Reading	1992-93	0	0	0	0	0	0	7	21	1	0	29
	1993-94	0	0	0	0	0	5	14	10	0	0	29
	1994-95	0	0	0	0	2	0	14	13	0	0	29
	1995-96	0	0	0	0	0	0	9	19	1	0	29
Science	1992-93	0	0	0	0	0	3	10	16	0	0	29
	1993-94	0	0	0	0	2	8	12	7	0	0	29
	1994-95	0	0	0	0	0	5	16	8	0	0	29
	1995-96	0	0	0	0	2	6	20	1	0	0	29
Social Studies	1992-93	0	0	0	0	0	4	13	12	0	0	29
	1993-94	0	0	0	0	1	5	16	7	0	0	29
	1994-95	0	0	0	0	0	3	13	13	0	0	29
	1995-96	0	0	0	0	0	11	13	5	0	0	29

TABLE 6-4
NOVICE VS. APPRENTICE, PROFICIENT, AND DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 4

Subject Area	Year	Biserial Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90+	
Mathematics	1992-93	0	0	0	0	0	7	8	11	3	0	29
	1993-94	0	0	0	0	2	3	9	13	2	0	29
	1994-95	0	0	0	0	0	5	10	13	1	0	29
	1995-96	0	0	0	0	0	6	11	7	5	0	29
Reading	1992-93	0	0	0	0	0	1	11	16	1	0	29
	1993-94	0	0	0	0	1	3	14	10	1	0	29
	1994-95	0	0	0	1	0	1	7	16	4	0	29
	1995-96	0	0	0	0	1	1	6	15	6	0	29
Science	1992-93	0	0	0	0	1	5	9	14	0	0	29
	1993-94	0	0	0	1	2	8	12	6	0	0	29
	1994-95	0	0	0	0	0	7	12	10	0	0	29
	1995-96	0	0	0	0	2	8	18	1	0	0	29
Social Studies	1992-93	0	0	0	0	3	4	13	9	0	0	29
	1993-94	0	0	0	1	1	7	13	7	0	0	29
	1994-95	0	0	0	0	0	3	13	13	0	0	29
	1995-96	0	0	0	0	1	10	11	5	2	0	29

TABLE 6-5
NOVICE AND APPRENTICE, VS. PROFICIENT AND DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 4

Subject Area	Year	Biserial Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90+	
Mathematics	1992-93	0	0	0	0	2	3	2	11	8	3	29
	1993-94	0	0	0	0	2	2	8	12	3	2	29
	1994-95	0	0	0	0	1	2	9	10	6	1	29
	1995-96	0	0	0	0	1	3	10	13	2	0	29
Reading	1992-93	0	0	0	0	0	1	10	13	4	1	29
	1993-94	0	0	0	0	2	5	14	8	0	0	29
	1994-95	0	0	0	0	1	1	18	9	0	0	29
	1995-96	0	0	0	0	0	1	12	14	2	0	29
Science	1992-93	0	0	0	1	3	4	10	6	5	0	29
	1993-94	0	0	0	0	1	6	8	12	2	0	29
	1994-95	0	0	0	0	0	6	14	8	1	0	29
	1995-96	0	0	0	0	2	9	13	5	0	0	29
Social Studies	1992-93	0	0	0	1	1	2	9	11	5	0	29
	1993-94	0	0	0	0	1	6	16	6	0	0	29
	1994-95	0	0	0	0	0	3	16	9	1	0	29
	1995-96	0	0	0	0	3	12	9	5	0	0	29

TABLE 6-6
NOVICE, APPRENTICE, AND PROFICIENT VS. DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 4

Subject Area	Year	Biserial Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90+	
Mathematics	1992-93	0	0	0	0	2	3	5	9	6	4	29
	1993-94	0	0	0	1	2	4	6	12	2	2	29
	1994-95	0	0	0	0	1	4	7	14	1	2	29
	1995-96	0	0	0	0	2	4	10	11	2	0	29
Reading	1992-93	0	0	0	0	0	2	14	5	6	2	29
	1993-94	0	0	0	3	2	5	7	6	6	0	29
	1994-95	0	0	0	0	3	9	13	3	1	0	29
	1995-96	0	0	0	0	1	5	14	7	2	0	29
Science	1992-93	0	0	0	3	1	6	0	5	2	0	17*
	1993-94	0	0	1	1	2	6	6	4	3	2	25*
	1994-95	0	0	0	0	1	10	6	8	4	0	29
	1995-96	0	1	0	2	3	8	5	7	2	1	29
Social Studies	1992-93	0	0	0	1	1	3	7	9	6	2	29
	1993-94	0	0	0	1	1	5	11	10	1	0	29
	1994-95	0	0	0	0	0	5	14	7	3	0	29
	1995-96	0	0	0	0	2	13	10	4	0	0	29

* Note that for some forms of the test, no students achieved a performance level of Distinguished. For these forms, no biserial correlation could be calculated.

TABLE 6-7
DISTRIBUTION OF ITEM SCORES FOR OPEN-RESPONSE ITEMS, GRADE 8

Subject Area	Year	Item Scores										N Items
		.00-.40	.41-.80	.81-1.20	1.21-1.60	1.61-2.00	2.01-2.40	2.41-2.80	2.81-3.20	3.21-3.60	3.61-4.00	
Mathematics	1992-93	0	0	4	12	11	0	2	0	0	0	29
	1993-94	2	5	8	7	7	0	0	0	0	0	29
	1994-95	0	4	12	10	2	2	0	0	0	0	30
	1995-96	0	0	13	8	8	1	0	0	0	0	30
Reading	1992-93	0	0	1	5	20	3	0	0	0	0	29
	1993-94	0	1	1	4	13	10	0	0	0	0	29
	1994-95	0	0	0	2	19	9	0	0	0	0	30
	1995-96	0	0	1	4	14	10	1	0	0	0	30
Science	1992-93	0	0	3	16	8	2	0	0	0	0	29
	1993-94	0	1	8	16	3	1	0	0	0	0	29
	1994-95	1	2	9	13	5	0	0	0	0	0	30
	1995-96	0	3	9	14	4	0	0	0	0	0	30
Social Studies*	1992-93	0	0	0	2	21	6	0	0	0	0	29
	1993-94	0	0	7	14	6	1	0	0	0	0	28
	1994-95	0	0	3	17	10	0	0	0	0	0	30
	1995-96	0	0	6	19	5	0	0	0	0	0	30
Arts and Humanities	1994-95	0	1	0	8	3	0	0	0	0	0	12
	1995-96	0	0	0	19	5	0	0	0	0	0	24
Practical Living	1994-95	0	0	1	3	7	1	0	0	0	0	12
	1995-96	0	0	0	5	17	2	0	0	0	0	24

*Note: the responses to one matrix-sampled social studies item were discarded, due to accidental repetition of a common item as a matrix item in 1993-94.

TABLE 6-8
DISTRIBUTION OF CORRECTED ITEM-TOTAL CORRELATIONS FOR OPEN-RESPONSE ITEMS, GRADE 8

Subject Area	Year	Correlation .00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	N Items
Mathematics	1992-93	0	0	0	0	2	20	7	0	0	0	29
	1993-94	0	0	0	1	5	21	2	0	0	0	29
	1994-95	0	0	0	0	2	20	8	0	0	0	30
	1995-96	0	0	0	0	3	19	8	0	0	0	30
Reading	1992-93	0	0	0	0	0	7	18	4	0	0	29
	1993-94	0	0	0	0	3	7	19	0	0	0	29
	1994-95	0	0	0	0	2	15	13	0	0	0	30
	1995-96	0	0	0	0	2	21	7	0	0	0	30
Science	1992-93	0	0	0	0	6	22	1	0	0	0	29
	1993-94	0	0	0	5	14	10	0	0	0	0	29
	1994-95	0	0	0	2	12	16	0	0	0	0	30
	1995-96	0	0	0	2	16	12	0	0	0	0	30
Social Studies*	1992-93	0	0	0	0	1	8	20	0	0	0	29
	1993-94	0	0	0	1	0	9	17	1	0	0	28
	1994-95	0	0	0	0	1	7	21	1	0	0	30
	1995-96	0	0	0	0	0	14	16	0	0	0	30
Arts and Humanities**	1994-95	0	0	0	2	10	0	0	0	0	0	12
	1995-96	0	0	0	1	8	3	0	0	0	0	24
Practical Living	1994-95	0	0	1	2	8	1	0	0	0	0	12
	1995-96	0	0	0	0	8	4	0	0	0	0	24

*Note: the responses to one matrix-sampled social studies item were discarded, due to accidental repetition of a common item as a matrix item in 1993-94.

** Note: only one correlation per form is presented for arts & humanities and practical living|vocational studies, as the corrected item-total correlation is between two items appearing on the same form.

TABLE 6-9
DISTRIBUTION OF ITEM-THETA CORRELATIONS FOR OPEN-RESPONSE ITEMS, GRADE 8

Subject Area	Year	Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	0	1	2	14	12	0	0	29
	1993-94	0	0	0	0	1	3	14	11	0	0	29
	1994-95	0	0	0	0	0	2	19	9	0	0	30
	1995-96	0	0	0	0	0	5	15	10	0	0	30
Reading	1992-93	0	0	0	0	0	0	6	17	6	0	29
	1993-94	0	0	0	0	1	3	3	20	2	0	29
	1994-95	0	0	0	0	1	1	10	17	1	0	30
	1995-96	0	0	0	0	0	3	19	8	0	0	30
Science	1992-93	0	0	0	0	0	3	14	12	0	0	29
	1993-94	0	0	0	0	1	7	13	8	0	0	29
	1994-95	0	0	0	1	2	6	18	3	0	0	30
	1995-96	0	0	0	0	2	12	13	3	0	0	30
Social Studies*	1992-93	0	0	0	0	0	1	6	20	2	0	29
	1993-94	0	0	0	0	1	1	8	15	3	0	28
	1994-95	0	0	0	0	1	0	10	16	3	0	30
	1995-96	0	0	0	0	0	1	13	16	0	0	30

* Note: the responses to one matrix-sampled social studies item were discarded, due to accidental repetition of a common item as a matrix item in 1993-94.

TABLE 6-10
NOVICE VS. APPRENTICE, PROFICIENT, DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 8

Subject Area	Year	Biserial Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	1	0	6	8	12	2	0	29
	1993-94	0	0	0	1	1	6	6	12	3	0	29
	1994-95	0	0	0	0	2	7	13	8	0	0	30
	1995-96	0	0	0	0	3	8	8	11	0	0	30
Reading	1992-93	0	0	0	0	0	2	6	18	3	0	29
	1993-94	0	0	0	1	3	2	6	14	3	0	29
	1994-95	0	0	0	0	1	4	9	11	5	0	30
	1995-96	0	0	0	0	0	5	12	12	1	0	30
Science	1992-93	0	0	0	0	1	4	15	9	0	0	29
	1993-94	0	0	0	0	2	6	13	7	1	0	29
	1994-95	0	0	0	1	3	6	13	7	0	0	30
	1995-96	0	0	0	0	4	7	14	5	0	0	30
Social Studies*	1992-93	0	0	0	0	0	3	14	11	1	0	29
	1993-94	0	0	0	0	1	3	10	13	1	0	28
	1994-95	0	0	0	0	1	0	11	15	3	0	30
	1995-96	0	0	0	0	0	1	13	16	0	0	30

* Note: the responses to one matrix-sampled social studies item were discarded, due to accidental repetition of a common item as a matrix item in 1993-94.

TABLE 6-11
NOVICE, APPRENTICE VS. PROFICIENT, DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 8

Subject Area	Year	Biserial Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	0	0	3	6	5	10	5	29
	1993-94	0	0	0	0	0	1	8	8	7	5	29
	1994-95	0	0	0	0	0	2	7	14	7	0	30
	1995-96	0	0	0	0	0	1	9	15	5	0	30
Reading	1992-93	0	0	0	0	0	1	3	10	12	3	29
	1993-94	0	0	0	0	0	4	3	17	5	0	29
	1994-95	0	0	0	0	1	1	10	15	3	0	30
	1995-96	0	0	0	0	0	3	18	9	0	0	30
Science	1992-93	0	0	0	0	0	2	5	13	7	2	29
	1993-94	0	0	0	1	1	5	12	8	2	0	29
	1994-95	0	0	0	0	2	1	12	11	2	2	30
	1995-96	0	0	0	0	2	8	12	3	4	1	30
Social Studies*	1992-93	0	0	0	0	1	0	0	8	16	4	29
	1993-94	0	0	0	0	1	1	2	14	10	0	28
	1994-95	0	0	0	0	1	0	8	18	3	0	30
	1995-96	0	0	0	0	0	1	10	18	1	0	30

* Note: the responses to one matrix-sampled social studies item were discarded, due to accidental repetition of a common item as a matrix item in 1993-94.

TABLE 6-12
NOVICE, APPRENTICE, PROFICIENT VS. DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 8

Subject Area	Year	Biserial Correlation											N Items
		-.09- -.00	.00- .09	.10- .19	.20- .29	.30- .39	.40- .49	.50- .59	.60- .69	.70- .79	.80- .89	.90- 1.00	
Mathematics	1992-93	0	0	0	0	1	0	6	8	12	2	0	29
	1993-94	0	0	0	0	0	0	3	5	7	6	8	29
	1994-95	0	0	0	0	0	0	0	5	13	8	4	30
	1995-96	0	0	0	0	0	0	1	7	9	13	0	30
Reading	1992-93	0	0	0	0	0	0	2	6	18	3	0	29
	1993-94	0	0	0	0	0	2	2	8	13	2	2	29
	1994-95	0	0	0	0	0	1	2	10	13	3	1	30
	1995-96	0	0	0	0	0	2	7	6	9	2	0	26*
Science	1992-93	0	0	0	0	0	1	4	15	9	0	0	29
	1993-94	0	0	0	0	0	2	2	4	4	3	2	17*
	1994-95	1	0	0	1	0	3	3	3	7	3	3	24*
	1995-96	0	0	0	0	0	5	1	0	0	0	2	8*
Social Studies**	1992-93	0	0	0	0	0	0	3	14	11	1	0	29
	1993-94	0	0	0	0	0	1	0	3	7	16	1	28
	1994-95	0	0	0	0	0	0	1	4	12	13	0	30
	1995-96	0	0	0	0	0	0	2	6	17	4	1	30

* Note that for some forms of the test, no students achieved a performance level of Distinguished. For these forms, no biserial correlation could be calculated.

** Note: the responses to one matrix-sampled social studies item were discarded, due to accidental repetition of a common item as a matrix item in 1993-94.

TABLE 6-13
DISTRIBUTION OF ITEM SCORES FOR OPEN-RESPONSE ITEMS, GRADE 11/12

Subject Area	Year	Item Scores										N Items
		.00-.40	.41-.80	.81-1.20	1.21-1.60	1.61-2.00	2.01-2.40	2.41-2.80	2.81-3.20	3.21-3.60	3.61-4.00	
Mathematics	1992-93	0	1	15	9	4	0	0	0	0	0	29
	1993-94	2	7	9	8	3	0	0	0	0	0	29
	1994-95	1	9	4	9	5	2	0	0	0	0	30
	1995-96	1	6	11	5	7	0	0	0	0	0	30
Reading	1992-93	0	0	1	11	13	4	0	0	0	0	29
	1993-94	0	1	0	6	14	8	0	0	0	0	29
	1994-95	0	0	0	7	18	5	0	0	0	0	30
	1995-96	0	1	11	17	1	0	0	0	0	0	30
Science	1992-93	0	0	4	22	3	0	0	0	0	0	29
	1993-94	0	1	8	12	7	1	0	0	0	0	29
	1994-95	0	1	6	14	8	1	0	0	0	0	30
	1995-96	0	0	8	16	5	1	0	0	0	0	30
Social Studies	1992-93	0	0	6	16	6	1	0	0	0	0	29
	1993-94	0	2	0	10	13	4	0	0	0	0	29
	1994-95	0	0	8	11	9	2	0	0	0	0	30
	1995-96	0	1	11	12	5	1	0	0	0	0	30
Arts and Humanities	1994-95	0	1	4	7	0	0	0	0	0	0	12
	1995-96	0	0	3	16	5	0	0	0	0	0	24
Practical Living	1994-95	0	0	2	3	7	0	0	0	0	0	12
	1995-96	0	0	2	6	16	0	0	0	0	0	24

TABLE 6-14
DISTRIBUTION OF CORRECTED ITEM-TOTAL CORRELATIONS FOR OPEN-RESPONSE ITEMS, GRADE 11/12

Subject Area	Year	Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	1	2	11	15	0	0	0	29
	1993-94	0	0	0	2	6	14	7	0	0	0	29
	1994-95	0	0	0	1	2	11	15	1	0	0	30
	1995-96	0	0	0	0	1	15	14	0	0	0	30
Reading	1992-93	0	0	0	0	1	9	18	1	0	0	29
	1993-94	0	0	0	2	2	12	13	0	0	0	29
	1994-95	0	0	0	0	1	15	14	0	0	0	30
	1995-96	0	0	0	0	1	19	10	0	0	0	30
Science	1992-93	0	0	0	0	2	17	10	0	0	0	29
	1993-94	0	0	1	0	10	12	6	0	0	0	29
	1994-95	0	0	0	1	7	14	8	0	0	0	30
	1995-96	0	0	0	0	6	23	1	0	0	0	30
Social Studies	1992-93	0	0	0	0	0	6	22	1	0	0	29
	1993-94	0	0	0	0	1	10	17	1	0	0	29
	1994-95	0	0	0	0	0	11	18	1	0	0	30
	1995-96	0	0	0	0	2	15	13	0	0	0	30
Arts and Humanities*	1994-95	0	0	0	1	5	6	0	0	0	0	12
	1995-96	0	0	2	2	5	3	0	0	0	0	24
Practical Living	1994-95	0	0	0	0	4	7	1	0	0	0	12
	1995-96	0	0	1	1	4	6	0	0	0	0	24

* Note: only one correlation per form is presented for arts & humanities and practical living\vocational studies, as the corrected item-total correlation is between two items appearing on the same form.

TABLE 6-15
DISTRIBUTION OF ITEM-THETA CORRELATIONS FOR OPEN-RESPONSE ITEMS, GRADE 11/12

Subject Area	Year	Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	0	0	3	10	16	0	0	29
	1993-94	0	0	0	0	3	5	10	10	1	0	29
	1994-95	0	0	0	1	1	4	14	9	1	0	30
	1995-96	0	0	0	0	0	2	16	12	0	0	30
Reading	1992-93	0	0	0	0	0	1	5	19	4	0	29
	1993-94	0	0	0	1	3	0	6	12	7	0	29
	1994-95	0	0	0	0	0	4	9	17	0	0	30
	1995-96	0	0	0	0	0	1	15	12	2	0	30
Science	1992-93	0	0	0	0	0	2	9	18	0	0	29
	1993-94	0	0	0	1	1	5	10	10	2	0	29
	1994-95	0	0	0	0	4	5	9	12	0	0	30
	1995-96	0	0	0	0	0	3	17	10	0	0	30
Social Studies	1992-93	0	0	0	0	0	2	4	17	6	0	29
	1993-94	0	0	0	0	0	3	9	16	1	0	29
	1994-95	0	0	0	0	0	1	13	16	0	0	30
	1995-96	0	0	0	0	0	2	11	17	0	0	30

TABLE 6-16
NOVICE VS. APPRENTICE, PROFICIENT, DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 11/12

Subject Area	Year	Biserial Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	0	3	6	6	14	0	0	29
	1993-94	0	0	0	0	4	8	9	7	1	0	29
	1994-95	0	0	0	0	0	5	9	15	1	0	30
	1995-96	0	0	0	1	3	8	15	3	0	0	30
Reading	1992-93	0	0	0	0	1	1	7	17	3	0	29
	1993-94	0	0	0	1	3	1	6	8	10	0	29
	1994-95	0	0	0	2	4	6	13	4	1	0	30
	1995-96	0	0	0	0	0	4	18	6	2	0	30
Science	1992-93	0	0	0	0	0	6	12	10	1	0	29
	1993-94	0	0	0	1	4	6	9	8	1	0	29
	1994-95	0	0	0	0	5	7	11	7	0	0	30
	1995-96	0	0	0	0	3	13	11	3	0	0	30
Social Studies	1992-93	0	0	0	0	0	5	3	16	5	0	29
	1993-94	0	0	0	0	1	5	10	13	0	0	29
	1994-95	0	0	0	0	0	5	10	15	0	0	30
	1995-96	0	0	0	0	0	5	12	13	0	0	30

TABLE 6-17
NOVICE, APPRENTICE VS. PROFICIENT, DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 11/12

Subject Area	Year	Biserial Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	1	1	2	2	7	7	9	29
	1993-94	0	0	0	0	2	3	3	10	7	4	29
	1994-95	0	0	0	1	0	1	5	14	7	2	30
	1995-96	0	0	0	0	0	0	7	11	9	3	30
Reading	1992-93	0	0	0	0	0	1	2	14	11	1	29
	1993-94	0	0	0	0	2	3	3	14	7	0	29
	1994-95	0	0	0	0	0	3	6	12	9	0	30
	1995-96	0	0	0	0	0	1	9	15	5	0	30
Science	1992-93	0	0	0	0	1	1	7	11	8	1	29
	1993-94	0	0	0	1	1	2	10	10	4	1	29
	1994-95	0	0	0	0	2	7	5	14	2	0	30
	1995-96	0	0	0	0	0	4	9	12	5	0	30
Social Studies	1992-93	0	0	0	1	0	1	1	12	9	5	29
	1993-94	0	0	0	0	0	1	6	15	7	0	29
	1994-95	0	0	0	0	0	1	8	13	8	0	30
	1995-96	0	0	0	0	0	2	6	17	5	0	30

TABLE 6-18
NOVICE, APPRENTICE, PROFICIENT VS. DISTINGUISHED
DISTRIBUTION OF BISERIAL CORRELATIONS FOR OPEN RESPONSE ITEMS, GRADE 11/12

Subject Area	Year	Biserial Correlation										N Items
		.00-.09	.10-.19	.20-.29	.30-.39	.40-.49	.50-.59	.60-.69	.70-.79	.80-.89	.90-1.00	
Mathematics	1992-93	0	0	0	1	0	1	2	6	8	11	29
	1993-94	0	0	0	0	2	2	4	9	6	6	29
	1994-95	0	0	0	1	0	1	3	8	8	9	30
	1995-96	0	0	0	0	0	2	5	6	6	11	30
Reading	1992-93	0	0	0	0	2	2	2	7	13	3	29
	1993-94	0	0	0	1	2	3	6	11	5	1	29
	1994-95	0	0	0	0	1	5	9	9	4	2	30
	1995-96	0	0	0	0	0	8	8	8	4	2	30
Science	1992-93	0	0	0	0	0	3	4	10	7	5	29
	1993-94	0	0	0	1	2	0	11	7	4	4	29
	1994-95	0	0	0	0	1	7	5	13	4	0	30
	1995-96	0	0	0	0	0	2	15	9	4	0	30
Social Studies	1992-93	0	0	0	1	0	3	5	8	4	8	29
	1993-94	0	0	0	0	0	0	6	13	6	4	29
	1994-95	0	0	0	0	0	3	3	7	10	7	30
	1995-96	0	0	0	0	0	2	3	11	11	3	30

SECTION III
ADMINISTRATION, SCORING, AND REPORTING

Section 3 deals with the operational heart of the KIRIS testing and accountability programs. Chapter 7 describes how all portions of the test were administered. Chapter 8 describes the scoring process for the test. Chapter 9 presents information about the equating process used to maintain standards across years and across test forms within years. Chapter 10 highlights the score reporting process - which reports were sent to schools and to districts and when were they sent. Chapter 11 presents the results for the second accountability cycle. Because of their complexity, scoring issues related to portfolios get a separate chapter of their own, Chapter 12.

CHAPTER 7
TEST ADMINISTRATION

DETERMINING STUDENTS FOR WHOM A SCHOOL IS ACCOUNTABLE

Beginning with the 1993-94 school year, schools were held accountable for open-response test scores of all students enrolled in the school on the first day of the on-demand testing window, as illustrated in Table 7-1. In that same year, schools were held accountable for a student's writing (and, where applicable, mathematics) portfolio score if the student was enrolled in the school on the day writing portfolios were due, as long as that student had been enrolled in any Kentucky public school for at least 100 days. This latter requirement, enrollment for 100 days, was removed for subsequent years. For performance events, schools were held accountable for students attending school on the day of testing; this policy did not change across the years.

TABLE 7-1
RULES FOR DETERMINING STUDENTS FOR WHOM A SCHOOL IS ACCOUNTABLE

Accountability Year	On-Demand Eligibility	Portfolio Eligibility
1992-93	Enrolled on 20th day of school year	Same as on-demand
1993-94	Grades 4 and 8 – enrolled on April 18, 1994 Grade 12 – enrolled on February 7, 1994	Enrolled on March 1, 1994, and enrolled in one or more Kentucky public schools for at least 100 days by March 1
1994-95	Grade 7/8 - April 17, 1995 Grade 11/12 - February 6, 1995	March 24, 1995 Math Portfolio - March 17, 1995
1995-96	Grades 4 and 8 - enrolled on April 15, 1996 Grade 11 - enrolled on February 12, 1996	March 22, 1996 (Same for Math Portfolio)

Any student for whom a school was accountable, but who was not tested or did not produce a portfolio, was assigned a score of *Novice*.

COLLECTING ENROLLMENT INFORMATION. School enrollment information was taken from information provided by students and teachers on the scannable Student Response Booklets. The Writing Portfolio Score Form had a lithocode tracking number that was assigned to the student once these forms were completed and scanned by Advanced Systems.

The enrollment information was verified by Advanced Systems in two ways for the On-Demand Assessment. The first was by comparing the scanned information to a "Student Accountability Roster" that schools were required to send to Advanced Systems. The roster contained the names of all students enrolled on the On-Demand Assessment "date of accountability". The second method was by cross-checking this information against the data provided by school on the *Principal's Certification of Proper Test Administration* form. This form documented the number of students enrolled in the school (on the date of accountability), how many were tested, how many were excluded (in accordance with KDE exclusion guidelines), and how many were not tested (in accordance with KDE guidelines).

For the Writing Portfolio Assessment -- schools returned a Writing Portfolio Score Form for all students enrolled on the portfolio date of accountability. This score form was attached to the Student Response Booklet and had the same lithocode tracking number as the Student Response Booklet. After the portfolios are scored by teachers, the score form was marked with a student Response Booklet with the matching lithocode to Advanced Systems. If a student was still enrolled on the on-demand date of accountability there was no issue with student enrollment, but if the student moved between the portfolio date of accountability and the on-demand date of accountability the school and student had an enrollment discrepancy. To account for this the *Principal's Certification of Proper Test Administration* form had a column for schools to account for students who moved between these dates. Advanced Systems then compared the portfolio file, on-demand file, and *Principal's Certification of Proper Test Administration* form to ensure the student was accounted for by the school. This same process worked in reverse for students who moved into a school between the accountability dates. Schools still completed a Writing Portfolio Score Form for the student but marked the form as moving into the school after the portfolio date of accountability.

EXCLUSIONS. The following exclusions were authorized by KDE for the On-Demand Assessment.

- a student eligible for language exemption according to Department of Education policy: students whose native language was not English and who had been enrolled in an English-speaking school for fewer than two years were exempt from testing.
- a foreign exchange student.
- a student with an identified disability who was eligible for an alternate portfolio according to criteria established by the Department of Education: students in this category were exempt from on-demand testing, performance events, and writing and mathematics portfolios. These students each developed an alternate portfolio that counted towards second accountability cycle baselines. Criteria for eligible students were detailed in the Program Advisory 92-OCAA-104 (February, 1993) and published in the *Kentucky Alternate Portfolio Teacher's Guide* and the *1992-93 District Assessment Coordinator Implementation Guidebook*.
- a student medically excluded from testing: a student could be excluded from testing for medical reasons if a signed doctor's statement was provided.

The same exclusions were authorized for the Writing Portfolio Assessment along with the addition of the 100-day Rule.

For 1993-94, students needed to be enrolled in a Kentucky public school (including treatment or detention centers, or home-bound instruction) for at least 100 instructional days prior to the portfolio completion date to be included for accountability (out-of-state, home, Department of Defense, and private schools are not considered Kentucky public schools). The term "instructional days" applies to those days that the school was in session and that students were scheduled for class work, thereby excluding professional development days, holidays, snow days, and weekends. Absences, suspensions, and expulsions were not reasons to adjust an individual student's number of instructional days.

MODIFICATIONS TO DATA FILES. If conflicts in data were noted during scanning or enrollment verification District Assessment Coordinators were notified to assist in the resolution of the conflict. If the data discrepancy was not able to be resolved at the district level the information was forwarded to KDE for resolution. Confirmed data changes were made to the master student data files by Advanced Systems. These changes were "flagged" within the master student data file by Advanced Systems if they affected the accountability of a student. Final rosters listing the names and scores of students included in accountability calculations accompanied the schools' accountability reports.

Students who were not exempt but did not attempt the on-demand test or portfolio were assigned to the *Novice* performance level in the accountability reports. The percentage of Kentucky's fourth, eighth and eleventh/twelfth graders assigned to "Novice" because they did not complete the tests dropped appreciably over the course of the second accountability cycle. Tables 7-2 through 7-4 show for grades 4, 8, and 11/12, respectively, the number of students classified as exempt from on-demand testing, the number classified as Novice because they did not attempt the tests, and the number who completed the on-demand tests.

The exempted 'other' category includes students who moved out of state (or to private schools) and were unavailable for testing. Due to differences in record keeping across years, a small number of additional students are included in the 'other' category in order to simplify the tables. Eligible students are noted above as 'not tested (novice)' in cases where no attempt has been made at any open-response question, or all open-response questions were given off-topic responses and no multiple-choice questions were attempted.

TABLE 7-2
GRADE 4 STUDENTS IN EACH ACCOUNTABILITY ELIGIBILITY CATEGORY

Accountability Eligibility Category	1992-93		1993-94		1994-95		1995-96	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Exempted: Language	45	0.1	69	0.1	72	0.1	71	0.1
Exempted: Special Education	176	0.3	136	0.3	194	0.4	324	0.7
Exempted: Medical	113	0.2	35	0.1	34	0.1	52	0.1
Exempted: Foreign Exchange	19	0.0	0	0.0	3	0.0	3	0.0
Exempted: Other	1868	3.6	270	0.5	574	1.1	306	0.6
Exempted: Total	2221	4.3	510	1.0	877	1.7	756	1.6
Eligible: Not Tested (*Novice*)	464	0.9	82	0.2	80	0.2	85	0.2
Eligible: Tested	49559	94.9	49704	98.8	49365	98.1	47641	98.3
Eligible: Total	50021	95.7	49786	99.0	49445	98.3	47726	98.4
Total	52242	100.0	50296	100.0	50322	100.0	48482	100.0

TABLE 7-3
GRADE 8 STUDENTS IN EACH ACCOUNTABILITY ELIGIBILITY CATEGORY

Accountability Eligibility Category	1992-93		1993-94		1994-95		1995-96	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Exempted: Language	23	0	24	0.0	49	0.1	65	0.1
Exempted: Special Education	246	0.5	157	0.3	153	0.3	399	0.8
Exempted: Medical	219	0.4	50	0.1	71	0.1	95	0.2
Exempted: Foreign Exchange	40	0.1	4	0.0	0	0.0	3	0.0
Exempted: Other	1667	3.1	420	0.8	489	0.9	412	0.8
Exempted: Total	2195	4.1	655	1.3	762	1.5	974	1.9
Eligible: Not Tested (*Novice*)	1313	2.5	320	0.6	298	0.6	343	0.7
Eligible: Tested	49537	93.4	51621	98.1	50915	98.0	50321	97.4
Eligible: Total	50850	95.9	51941	98.8	51213	98.5	50664	98.1
Total	53045	100.0	52596	100.0	51975	100.0	51638	100.0

TABLE 7-4
GRADE 12/11 STUDENTS IN EACH ACCOUNTABILITY ELIGIBILITY CATEGORY

Accountability Eligibility Category	1992-93		1993-94		1994-95		1995-96	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Exempted: Language	12	0.0	14	0.0	26	0.1	42	0.1
Exempted: Special Education	86	0.2	116	0.3	105	0.2	161	0.4
Exempted: Medical	327	0.6	78	0.2	43	0.1	107	0.3
Exempted: Foreign Exchange	66	0.1	179	0.5	76	0.2	127	0.3
Exempted: Other	1199	2.9	242	0.6	148	0.4	132	0.3
Exempted: Total	1690	4.3	629	1.7	398	0.9	569	1.4
Eligible: Not Tested (*Novice*)	1827	4.6	164	0.4	399	0.9	346	0.8
Eligible: Tested	36224	91.1	36597	97.9	41298	98.1	41009	97.8
Eligible: Total	38051	95.7	36761	98.3	41697	99.1	41355	98.6
Total	39741	100.0	37390	100.0	42095	100.0	41924	100.0

ADMINISTRATION OF OPEN-RESPONSE TESTS

GENERAL PROCEDURES. Table 7-5 presents the test dates for each year of the first accountability cycle.

TABLE 7-5
TESTING WINDOWS*

Grade	Test Type	Year			
		1992-93	1993-94	1994-95	1995-96
4	On-demand	April 19 - 30	April 19 - 30	April 17 - May 28	April 15 - May 3
	Performance Events	April 7 - 23	April 11 - 29	April 16 - 28	April 22 - May 10
8	On-demand	April 19 - 30	April 19 - 30	April 17 - 28	April 15 - May 3
	Performance Events	April 19 - May 7	April 25 - May 13	April 27 - May 12	February 12 - March 1
11/12	On-demand	February 8 - 19	February 8 - 19	February 6 - 20	February 12 - March 1
	Performance Events	March 1 - 12	March 7 - 18	February 27 - March 10	March 8 - 29

*Not including make-up testing

As an example, for the 1995-96 school year, testing of all grade 11 students in Kentucky public schools occurred between February 12 and March 1, 1996. Testing for grades 4 and 8 took place April 15 through May 1. Within these testing windows, schools were allowed to set up their own specific testing schedules. Coordinators' and administrators' manuals served as a guide to administration. The manuals detailed timing requirements, directions for students, and other considerations for administering the tests and handling materials. *Appropriate Assessment Practices,* a document outlining appropriate behaviors for school personnel during testing, was included in the test administration manual; a copy of this document is available from the Kentucky Department of Education.

Eight subject areas were assessed using open-response tests: reading, mathematics, science, social studies, Arts and humanities, practical living, vocational studies and writing prompts. Each subject was administered in one 90-minute test session. Test administration procedures in 1993-94 provide students with as much additional time as needed as long as constructive progress is being made toward completion of the test. Additional time was scheduled directly after the initial test session for the subject area.

Students received a test booklet and a separate scannable answer document in which they recorded their answers to all questions, including open-response questions. (A sample answer document is included in Appendix 7-1 of the KIRIS Accountability Cycle I Technical Manual.) Twelve forms of the test were created; students were randomly given one form that they used for all sections of the test. In other words, if they received Form 1 for the reading test, they completed Form 1 for other subject areas as well.

Each school principal completed a *Principal's Certification of Proper Test Administration* to confirm enrollment and testing figures. The principal also certified on this form that the testing was conducted in accordance with the instructions provided in the manuals and that all materials were handled in a secure manner. A copy of the form is presented in Appendix 7-1, along with a similar form completed by District Assessment Coordinators.

SHIPPING AND RECEIVING PROCEDURES. All materials were packaged by the school and shipped to District Assessment Coordinators. The materials were sent to district offices approximately one week prior to the beginning of test administration. All district and school personnel were informed that the materials were secure and that all materials had to be returned to Advanced Systems at the completion of testing.

District Assessment Coordinators were instructed to package all materials from all schools in their districts within one week after the completion of testing. The shipping company picked up packaged materials from every district office.

As packages of answer sheets were received at Advanced Systems, they were collected in a prescribed location for check-in. Packages were opened one at a time, answer sheets were counted, and the count was compared with information provided by the school principal on the *Principal's Certification of Proper Test Administration*. The count was entered in a log, and any discrepancies were noted both on the log and on the principal's form.

To prepare the answer sheets for hand-scoring, the cover page, was separated from the open-response pages. Cover pages were scanned and the open-response pages left in the booklets were separated by form. These were then shuffled to ensure that multiple scorers would evaluate open responses from each school.

Test booklets and other material were checked for stray answer documents or other forms mistakenly shipped with them. The test booklets were then destroyed.

ADMINISTRATION OF PERFORMANCE EVENTS

The administration of the performance events testing in the second Accountability Cycle was managed by Advanced Systems' office in Louisville, Kentucky. In addition to the Regional Director and the Supervisor of Project Support Services, the staff included four regional events managers and a cadre of 73-80 (depending on the year) specially trained facilitators who traveled to each school to administer the event tasks. This section describes general procedures; significant variations for any one year are included in the narrative.

Tentative testing dates for each school were sent to principals by mid-October, requests for changes were accepted and accommodated, and final test date confirmations were mailed before Thanksgiving. Principals were asked to designate a school contact who would serve as ASME's point of communication for all future interchange with regard to performance events.

After scheduling was completed, events managers were responsible for coordinating all facets of performance events administration in a specific geographic area of the state. Their duties included:

- individual phone consultations with designated contacts at each school to confirm schedules and plan individual testing points
- checking the list of schools included for testing for correctness and completeness
- recruiting, interviewing and hiring test facilitators (each manager hired 18 - 22 facilitators from his/her region, to cover schools within that region)
- developing facilitator testing schedules
- monitoring and trouble-shooting throughout the weeks of events administration (examples: replacing or substituting for facilitators who were unable to meet prescribed schedules; re-scheduling test dates canceled due to unforeseen circumstances; responding to school/district calls for information/clarification; record-keeping, etc.).

Additional temporary personnel were employed by Advanced Systems' Kentucky office during the events administration to fill roles including clerical support and light labor required to pack performance events testing materials.

Applicants for the facilitator positions were recruited through the following mechanisms:

- position announcements were mailed to local chapters of the Kentucky Retired Teachers Association for distribution to members
- notices were sent to placement offices of Kentucky state colleges and universities
- the Kentucky Department of Employment Services listed the position in its statewide job bank and authorized Advanced Systems' staff to conduct interviews at the 14 sites across the state
- advertisements were placed in the Courier-Journal (state edition), the Owensboro Messenger-Inquirer, the Lexington Herald-Leader, the Paducah Sun Democrat, the Ohio County Messenger, and the Louisville Defender
- notices were sent to Urban League offices
- individual contacts with school boards, public school educators, colleges and universities, and community groups.

Managers screened hundreds of applications and annually conducted approximately 250 interviews to hire the facilitators. Hiring decisions were based on applicant qualifications in accordance with the requirements set forth in the position announcement; preference was given to applicants with bachelor's degrees and/or experience in school classrooms.

Table 7-6 profiles the facilitators employed to administer the performance events.

TABLE 7-6
PROFILE OF PERFORMANCE EVENTS FACILITATOR QUALIFICATIONS AND DEMOGRAPHICS

		1992-93	1993-94	1994-95	1995-96
Education	Degrees beyond the baccalaureate	23	34	45	50
	Bachelor's degree	42	36	36	32
	Associate's degree	9	2	1	3
	Some college study	0	1	1	1
Experience	School Principal/Administrator	5	3	8	14
	Teacher	39	27	33	45
	School Aide	17	22	15	4
	Other/Non-school	13	21	27	23
Demographics	Male	21	23	21	21
	Female	53	50	62	65
	African-American	3	2	0	0
	White	71	71	83	86

TRAINING OF FACILITATORS. Training of facilitators was conducted by Advanced Systems' Kentucky staff. Staff members from the Kentucky Department of Education typically attended some or all of each training session. For each year of the administration, training was conducted in two intensive two-day sessions in Louisville.

Annual training sessions included an overview of the Kentucky Education Reform Act; an introduction to the philosophy and purposes of performance assessment; extensive practice with the different assessment tasks, kits, and administration directions for the grade level; briefings on school protocol; and management of logistical issues such as travel, packing, and the collection and return of student response forms. Concerns that could affect test administration and its place in the accountability measure were also examined and discussed. These issues included: testing special needs students; random assignment of students to groups; the role of observers/teachers in the testing sessions; and appropriate responses to difficult, unpredictable or attitudinal problems in the schools. At the conclusion of each training session, facilitators received the performance events kits, materials, and forms they needed to conduct testing.

ADMINISTRATION PROCEDURES. Advanced Systems' facilitators delivered materials necessary for performance test administration, including the Student Response Form and the unique supplies and manipulatives required for each task, on the scheduled day of testing for each school. Facilitators worked with designated school contacts to conduct testing following operational directions and policy. Facilitators set up materials for each testing session, randomly assigned students in each session to a task group, gave general directions and timed the tasks, collected all testing materials, and prepared "Session Reports" and "Participation Rate" forms. At the conclusion of testing, facilitators collected all student products and returned them to the New Hampshire office of Advanced Systems for scoring.

Performance events were administered to all students in attendance on the day of testing in grades 4, 8, and 11 or 12 (in accordance with State Board decisions, the high school testing was moved from the 12th grade to the 11th grade for the 1994-95 school year and beyond) at each school designated A-1, a classification indicating full accountability as a Kentucky school. (Beginning with the 1994-95 school year, schools with fewer than nine non-exempt students present on the testing day were exempted from performance events. This was done to ensure that enough tasks could be administered. Scores for schools with fewer than nine students were computed from the students' performance on the open-response tests.) The events at each grade level included math, science, social studies. Arts and humanities, practical living and vocational studies were included beginning with the 1992-1993 school year. The tasks, which are summarized in Appendix 7-2, were typically designed for groups of three or four students. In keeping with KERA, schools were to have virtually all enrolled students participating in the events; specific exclusions were detailed in the "Commissioner's Program Advisories" of January 31, 1992 and February 9, 1993.

In general, students participated in only one event and in only one testing session. To minimize the sharing of information between students tested early in the day and those tested later, the testing at grades 8 and 11 (or 12, depending on the year) was arranged so that all students who completed a particular event did so at the same time. Grade 4 students were generally tested with their self-contained classroom, so the potential for them to share information was minimized and those administrations were arranged so that half of the events were taken by each classroom. After re-evaluating the potential for older students to share information, the events administration for the 8th and 11th grade students was altered in the final year (1995-96) to mirror the arrangement used at the 4th grade. This allowed Advanced Systems to streamline the amount of materials facilitators had to carry, and simplified the facilitators' training requirements.

Small schools (those with fewer than 36 students at an accountability grade level) were treated differently for test administrations. For the first two years of the second accountability cycle (1992-93, and 1993-94), students took two tasks, in two different testing sessions. Facilitators administered as many different tasks as was possible, given the limited number of students. (No tasks were administered twice unless every task had been administered at least once.) This procedure was altered for the final two years (1994-95 and 1995-96) of administration, after the Kentucky Department of Education determined that it would be better to cover fewer content areas per school in greater depth, rather than more content area with less depth. Therefore, for those years of the administration, students in small schools still participated in two events but when a school had from nine to 18 students, only arts and humanities, and practical living/vocational studies tasks were administered. (School data from the other content areas were derived from student performance on the open-response testing. Because students took fewer open-response items in arts and humanities, and practical living/vocational studies, it was determined that those were the areas which should be prioritized for events administration.) Schools with fewer than nine non-exempt students were exempted from performance events testing; their accountability scores were derived entirely from the open-response testing.

Schools were asked to provide a room or rooms in which testing could be conducted, a teacher or other member of the instructional staff to help maintain order and observe the process, and specific standard materials (equipment and resources typically found in schools) needed for the tasks. Principals had the option of allowing additional members of their instructional staff to observe each testing session (two was the recommended maximum number). Principals were asked to ensure that students who used special equipment or resources (including aides) in their regular instructional programs (as specified in an IEP or 504 plan) had those resources available during testing, and to identify children with special physical needs so that facilitators could make appropriate task assignments.

MAKE-UP SESSIONS. Make-up sessions for an entire school were held only if unforeseen circumstances (e.g. school closings due to inclement weather) prevented testing on the originally scheduled date. There was no make-up testing for students who were not present on the test day. Schools were held accountable for the performance of students present in school on the day of testing, and were expected to have a participation rate of at least 90% of the students enrolled in the school on the day of testing. Using enrollment data supplied by principals, facilitators calculated participation rates before leaving the school at the end of each testing day, leaving that information with principals. KDE was sent a copy of the participation rate forms for those schools with less than 90% participation; these schools were required to provide an explanation to the Department of Education.

ADMINISTRATION OF PORTFOLIOS

KIRIS writing portfolios were administered in the second accountability cycle following the model established in 1991-92. A detailed description of the structure and process of training for development and scoring portfolios can be found in the *1991-92 Technical Report* and in *The Writing Portfolio Development and Scoring Process*, distributed in March, 1993. In 1992-93, mathematics portfolios were added to the KIRIS assessment in the accountability grades. The entire mathematics portfolio process followed the structure and procedures established for writing portfolios. There were no major changes over the course of the second accountability cycle from the math portfolio process implemented in 1992-93.

TEACHER TRAINING. The training model for portfolio development and scoring followed the pyramidal region/cluster organization established for 1991-92. The state was divided into eight regions, with each region assigned a Regional Coordinator for each accountability grade. Regional Coordinators participated as training Cluster Leaders in their regions and served as members of their respective advisory committees. Cluster Leaders are the teachers who trained all participating teachers, in "clusters" of 20–25, to develop and score portfolios. Most initial training focused on developing and implementing portfolios in the classroom. Later training tended to focus on scoring portfolios. For writing portfolios, District Assessment Coordinators (DACs) in local districts selected Cluster Leaders. For mathematics portfolios, Cluster Leaders made application to, and were selected by, the Department of Education.

TRAINING FOR DEVELOPMENT AND IMPLEMENTATION. The first phase of training addressed the development of portfolios, fulfillment of requirements, and state guidelines for the generation of student work. For both writing and mathematics, Regional Coordinators were the first to receive this intensive training. Regional Coordinators then provided the same training for Cluster Leaders. Advanced Systems and Kentucky Department of Education consultants, assisted by Regional Coordinators, trained the Cluster Leaders. Cluster Leaders then returned to their local districts to provide the same training to all classroom teachers involved in the development of portfolios. KDE and Advanced Systems' consultants also conducted training that was telecast by Kentucky Educational Television (KET). Districts were encouraged to record the telecasts, or could order the videocassette directly from KET. The following materials were used during the first phase of training.

Writing Portfolios:

- *The Kentucky Writing Portfolio Teacher Handbook*
- "The Kentucky Writing Portfolio Holistic Scoring Guide"
- "Assessment Update: Portfolio Panel," aired October 7, 1992 (videocassette available)

Mathematics Portfolios:

- *The Kentucky Mathematics Portfolio Teacher's Guide*
- Sample Portfolios
- *Kentucky's Preliminary Curriculum Framework* (Goal II, pp. 179-208)
- "Assessment Update 5: Getting Ready for Math Portfolios," aired May 6, 1992 (videocassette available)
- "Assessment Update: Getting Started with Math Portfolios" (cluster training), aired September 2, 1992 (videocassette available)
- "Assessment Update: Portfolio Panel," aired October 7, 1992 (videocassette available)

TRAINING FOR SCORING. Teachers were trained on the standards and procedures for scoring writing and/or mathematics portfolios. This was accomplished through the same training procedures as for development and implementation. Telecasts were also used for this part of mathematics portfolio training. The following materials were used during this second phase of training.

Writing Portfolios:

- *The Kentucky Writing Portfolio Teacher Handbook*
- "The Kentucky Writing Portfolio Holistic Scoring Guide"
- *The 1993-94 Kentucky Writing Portfolio Scoring Training Manual*

Mathematics Portfolios:

- *The Kentucky Mathematics Portfolio Teacher's Guide*
- "The Kentucky Mathematics Portfolio Holistic Scoring Guide"
- *KIRIS Mathematics Portfolio Assessment Scoring Training Manual*
- "Assessment Update: Timeline for Scoring Math Portfolios" (cluster training), aired March 3, 1993 (videocassette available)
- "Mathematics Portfolios Scoring Training," Overview and Grades 4, 8, and 12, aired June 16, 1993 (four separate videocassettes available)

ADDITIONAL TRAINING AND MATERIALS. In addition to the two established phases of training for all participating teachers in the state, additional optional training and supplemental materials, described below, were offered throughout the 1992-93 school year. In all regions of the state, KDE and Advanced Systems conducted optional guided scoring sessions to provide a supportive scoring environment and a forum for discussion of portfolio issues and scoring standards for both mathematics and writing portfolios.

With regard to Writing Portfolios, a question-and-answer brochure, addressing some of the most commonly-asked questions from teachers, was compiled by the Department of Education and Advanced Systems consultants and distributed to writing portfolio Cluster Leaders. Also, the Writing Advisory Committee met to score a stratified random sample of portfolios and identify a set of Consensus Portfolios reflecting state standards. These were used to train and qualify scorers who participated in an important series of portfolio rescoring sessions, some of which are described in Chapter 12. These studies included scoring analysis, whereby school-level scoring feedback was provided to approximately half the schools based on 1993-94 writing portfolio scores, and to nearly all remaining schools based on 1994-95 scores. (A small number of schools declined to participate in scoring analysis.) All schools were eligible for inclusion in an audit of 1995-96 scores.

With regard to Mathematics Portfolios, all teachers of mathematics at an accountability grade were provided with addenda to the *Kentucky Mathematics Portfolio Teacher's Guide* that consisted of the following: "Guidelines for the Generation of Student Work;" two additional sample portfolio tasks per accountability grade; the "Self-Assessment Conference Form;" a bibliography; and a question-and-answer document that addressed the most frequently asked questions across the state. The addenda materials were created and/or reviewed by the Regional Coordinators at an implementation issues meeting held in late fall. Concurrent with the March, 1993, dissemination of this addenda, regional Cluster Leader Meetings were held to review the new materials and clarify other portfolio issues. A KET telecast was also conducted to assist Cluster Leaders in sharing the new materials with their cluster members.

ALTERNATE PORTFOLIO. Alternate portfolio administration followed a process similar to the writing and mathematics portfolios, including training for development and implementation. Alternate portfolios differ dramatically from the other portfolios with respect to determining student eligibility and portfolio scoring.

The alternate portfolio program is for students with disabilities who meet the criteria for eligibility specified in the Program Advisory OCAA-93-104. This is a small population of students whose programs in school are significantly different from those of other students. These programs are tailored to the Individual Education Program (IEP). Instead of participating in on-demand testing and writing portfolio assessment, these students submit an alternate portfolio for inclusion in a school's accountability index at 4th grade, 8th grade and their last year of school.

The alternate portfolio is scored twice at the regional level by trained teachers and administrators from other districts. Only teachers trained to score the alternate portfolio during each school year may score during the regional scoring sessions. Should the score assigned to a portfolio at the regional scoring session differ by 2 or more performance levels from the original score, the portfolio is sent to the state level scoring session. A committee composed of regional leaders, district level leaders, and alternate portfolio staff score these discrepant portfolios. In these cases, the third score is used as the accountability score.

Documents used in the development of portfolio scoring training were:

- KIRIS Alternate Portfolio Teachers Guide
- KIRIS Alternate Portfolio Addendum
- KIRIS Alternate Portfolio Scoring Manual

ASSESSMENT OF NONCOGNITIVE INDICATORS

In addition to the cognitive indicators used to measure school success, KERA mandates that noncognitive indicators be factored into the accountability index. A subcontractor for the assessment, WestEd, collects and analyzes the noncognitive data. The noncognitive indicators represented sixteen percent of the overall school accountability index in the second accountability cycle, and were composed of several indicators, not all of which are used at each grade level.

- **Attendance Rate**, defined as the ratio of aggregate attendance to the sum of aggregate attendance and aggregate absence (used at all grade levels);
- **Retention Rate**, defined as the ratio of number retained to total membership (used at grades 4 to 12);
- **Dropout Rate**, defined as the ratio of number of dropouts to total membership (used at grades 7 to 12 only); and
- **Successful Transition Rate**, defined as the ratio of number of "successful" high school graduates to the total number of graduates. As detailed in 1991 by the Department of Education and published in *Guidelines for Transition to Adult Life*, "successful transition to adult life" is identified by such criteria as full-time enrollment in post-secondary school, employment in a non-temporary position at least 30 hours per week, active military status, or a combination of these situations totaling at least 30 hours per week. Transition is assessed by school administrators who must verify "successful" transition with either the student or a member of the student's family (used in grade 12 only).

LAGGING DATA TO ACCOMMODATE TIMELY REPORTING OF SCORES. To facilitate reporting accountability index scores to schools in October of the year following test administration, it was necessary, beginning with 1994-95 test results, to lag the noncognitive data by one year. That is, the school accountability index for 1994-95 test scores was calculated using noncognitive indicators based on 1993-94 attendance, dropout, retention and successful transition rate figures. Scores for 1994-95 and 1995-96 were lagged in this regard because successful transition rate data from a given year were not available by the mid-September deadline necessary for inclusion in scores reported in October, whereas other non-cognitive indicators were available only in preliminary form, if at all. Thus, in the second accountability cycle, school year 1992-93 noncognitive data were used for 1992-93 accountability indices, school year 1993-94 noncognitive data were used for the 1993-94 and 1994-95 accountability indices, and school year 1994-95 noncognitive data were used for the 1995-96 accountability indices.

ACCURACY OF NONCOGNITIVE DATA. Noncognitive data from 1991-92 and 1992-93 were examined in a study of the accuracy of the noncognitive indicators and the procedures used to collect the data. (A full discussion of reliability issues for KIRIS noncognitive indicators can be found in Appendix F of the 1992-93 Technical Report.) In general, the analyses focused on:

- the thoroughness of the data-collection procedures including check-in and auditing strategies;
- stability within years, as indicated by modifications of submitted data following review by school and district staff of their original submissions;
- stability across years, as indicated by comparisons of 1992-93 data to previous year's data; and
- interpretability of the data across indicators at various grade and statewide levels

While these analyses found the procedures used to collect and report KERA noncognitive indicators to be thorough and reliable, they suggested that KDE expand the number of options for districts to submit data, emphasizing methods that take advantage of improving local technology capabilities, and at the same time standardize noncognitive data-collection procedures and timelines and eliminate duplication of effort. In addition, findings suggested that KDE should increase the training of local staff on the modified definition of dropouts through memoranda, meetings, and other techniques, as well as consider a pilot study examining the accuracy of local data-collection procedures across the indicators.

CHAPTER 8
SCORING

OPEN-RESPONSE QUESTIONS AND ON-DEMAND WRITING

The 1994–95 and 1995–96 KIRIS open-response questions, on-demand writing samples, and performance events (for the 1994–95 administration only) at grades 4, 8 and 11 required hand-scoring by Advanced Systems personnel. Portfolios in writing and mathematics as well as alternate portfolio were hand-scored by Kentucky teachers. While the processes of selecting and training scorers, reading and scoring papers, and monitoring scoring remained very similar to those carried out in previous years, these procedures are described below in detail. Modifications in training materials for scorers and teachers are described as well. (For data on 1992-93 and 1993-94 scoring, consult the KIRIS Accountability Cycle I Technical Manual.)

STAFFING. Levels of staffing are listed in Table 8–1. The table also shows the percentages of scorers at each grade level who participated in a previous year's scoring (repeat scorers), as well as the number of training leaders. The comparatively low percentage of repeat scorers in 1995-96, particularly at grade 8, is due in part to opening a second scoring center to handle several forms of the test for grade 4 and 8.

QUALIFICATIONS. Table 8–2 shows education level and demographic information for scorers in each testing year.

SCORER TRAINING. The training of scoring staff was completed in two phases: training of the chief readers and training leaders, followed by training of the scorers.

The chief reader and the training leaders for open-response questions in each content area met with the Advanced Systems' test developer responsible for that domain. The developer, as a member of the Kentucky advisory committee, presented the Kentucky objectives, content guidelines, standards, and background information necessary to understand the objectives being measured. Each group also reviewed the framework of the scoring guide and the language pertinent to the standard.

After this introduction, the combined group read hundreds of student responses and selected benchmark exemplars — papers which typify each score point in the scoring guide. Once the benchmarks were established, the training leaders continued the preparation. They identified a second set of benchmark papers, similar to the first training pack, and additional examples

TABLE 8–1
NUMBER OF SCORERS AND TRAINING LEADERS AT EACH GRADE

Grade	1994–95			1995–96		
	% Repeat Scorers	Scorers	Training Leaders	% Repeat Scorers	Scorers	Training Leaders
4	87	203	14	63	267	21
8	83	263	17	9	221	20
12	67	172	14	73	167	25

TABLE 8–2
PROFILE OF SCORER QUALIFICATIONS AND DEMOGRAPHICS

Background		Number of Scorers					
		1994–95			1995–96		
		Grade 4	Grade 8	Grade 12	Grade 4	Grade 8	Grade 12
Education	Degrees beyond the Baccalaureate	35	52	25	52	18	42
	Bachelor's Degree	110	156	119	159	78	104
	Associate's Degree	4	5	4	9	19	8
	Two-year college study or equivalent and previous scoring experience	54	50	24	47	106	13
Demographics	Male	76	97	70	104	91	71
	Female	127	166	102	163	130	96
	Black	2	4	2	3	6	3
	White	201*	259*	170*	261	215*	161
	Other	0	0	0	3	0	3

*Data collection did not include an "other" category.

of student responses that represented a range within each score point. Throughout this process, development staff was available to discuss concerns presented by the training leaders and answer any questions that they might have. Before training packs were reproduced, the training leader met with the developer for a final review of the training materials.

Advanced Systems' development staff observed the initial sessions when the training leaders presented the standards to the scorers and were available to provide additional clarification if needed. The training leader then completed the training independently. The training leader and developer consulted as needed throughout scoring; at the end of the project, development staff and leaders met to share information about the process and to offer suggestions and comments for future improvement. Throughout the scoring period, training leaders read student responses as they checked the reliability of each scorer. As they read, training leaders also selected recalibration papers that were reviewed with scorers at the beginning of each day to assure that drift from scoring standards did not occur.

Scorers for each content area were selected for their content expertise and were trained by the training leaders. The training leaders first presented background information and an explanation of the scoring guide. The first set of training papers — the benchmark training pack — was used to clarify the language of the scoring guide; each score point was illustrated by an exemplar. This pack became the reference set used throughout scoring. Scorers were instructed to review the language of the guide regularly as they read actual student responses.

The first training pack was similar to the benchmark pack, but papers were in random rather than sequential order by score point. A second training pack was designed to teach scorers how to identify a range within each score point. After discussing the papers in each pack, scorers were asked to assign scores independently to another set of papers. The scores were compared to those assigned by the training leaders and developers. The final training pack consisted of unique student responses to the question or writing prompt.

As a final qualifying step, scorers were instructed to score six papers. The training leaders checked each person's responses for accuracy; those who achieved success on four of the papers in assigning the appropriate score point began to score actual papers. Those who needed further training worked with the training leader until they were able to achieve the two-thirds rate of agreement required to qualify. All scorers hired for the project qualified in this process.

Training papers used in 1992-93 and 1993-94 included a benchmark set and a training pack. The benchmark pack expanded on the language of the scoring guide and provided exemplar responses to illustrate each score point. The training pack included several student samples in random score order, some clearly similar to benchmark samples, some different. Scorers' work on this set was compared to those scores assigned by the training leaders and developers to determine whether they were ready to begin actual scoring.

Training for the scoring of writing was similar in all ways to that of the open-response questions except for the initial step. The chief reader and training leaders for writing met with the Advanced Systems writing consultant and the Kentucky writing committee to establish the standards and select the benchmark papers. Other procedures were the same for the rest of the preparation and training process.

CONSISTENCY OF SCORING. Scoring of open-response tests was monitored in two ways. The first was ongoing, as training leaders constantly moved from scorer to scorer, re-reading samples of each scorer's work. Each training leader read approximately two papers per scorer each morning and afternoon.

The second monitoring procedure was a rescoring of all readers, including the training leaders. Two percent of the total responses were rescored in this way. Tables 8-3 and 8-4 document the percentage of exact matches between item-level scores assigned by separate scorers. Table 8-5 presents correlations of student-level scores assigned by separate scorers to open-response questions. These statistics indicate a high degree of consistency across scorers. Less than one percent of the scores subject to rescoring were found to be discrepant by more than one point, according to the verification data collected. (No adjustments were made, however, to discrepant scores.) This finding is similar to figures for 1992-93 and 1993-94, as presented in the KIRIS Accountability Cycle I Technical Manual.

PERFORMANCE EVENTS

When the team of scorers completed the scoring of grade 8 open-response questions, the final set of open-response materials to be scored, many were retained to score the 1994-95 KIRIS performance events for all grade levels. The training procedures and score monitoring by the training leader were the same as for the open-response tests. That is, Advanced Systems' test development staff and content advisory committees were responsible for writing scoring guides for each task and then Advanced Systems staff worked with the training leaders to select student responses that illustrated score points in the guides. The student responses were assembled into packets that were used to train the group of scorers.

TABLE 8-3
INTER-RATER RELIABILITY IN SCORING OF OPEN-RESPONSE QUESTIONS

Grade	Percentage of Exact Agreement							
	1994-1995				1995-1996			
	4	8	11	Total	4	8	11	Total
Reading	85.7	86.2	84.8	85.6	89.5	90.3	88.6	89.5
Math	95.3	97.1	93.6	95.5	96.4	98.5	97.9	97.6
Science	87.3	85.9	86.2	86.4	86.3	82.5	80.9	82.9
Social Studies	85.0	85.2	84.0	84.8	83.3	80.0	79.3	80.8
Arts and Humanities	87.7	87.6	87.1	87.5	90.1	87.3	89.8	88.8
Practical Living	89.3	87.3	88.7	88.3	90.2	88.7	88.5	89.5
Total	87.8	87.9	86.8	87.5	88.7	86.6	85.5	86.9

TABLE 8-4
INTER-RATER RELIABILITY IN SCORING OF OPEN-RESPONSE QUESTIONS

Grade	Percentage of Within 1 Score Point							
	1994-1995				1995-1996			
	4	8	11	Total	4	8	11	Total
Reading	99.8	99.9	99.8	99.8	99.8	99.9	99.8	99.8
Math	99.9	99.9	99.8	99.9	99.4	99.7	99.7	99.6
Science	99.8	99.8	99.7	99.8	99.4	99.0	99.1	99.2
Social Studies	99.9	99.9	99.9	99.9	99.8	99.5	99.6	99.6
Arts and Humanities	99.8	99.9	99.7	99.8	99.9	99.8	99.9	99.8
Practical Living	100.0	100.0	99.9	100.0	99.9	99.9	99.9	99.9
Total	99.9	99.9	99.8	99.9	99.7	99.5	99.6	99.6

TABLE 8-5
STUDENT-LEVEL INTER-RATER CORRELATIONS

Subject	1994-1995			1995-1996		
	Grade 4	Grade 8	Grade 12	Grade 4	Grade 8	Grade 12
Reading	.92	.88	.90	.93	.93	.92
Math	.98	.99	.98	.97	.99	.98
Science	.92	.91	.92	.91	.91	.90
Social Studies	.91	.91	.92	.90	.87	.88
Arts and Humanities	.91	.91	.90	.94	.88	.91
Practical Living	.92	.91	.91	.92	.91	.90

Scores for performance events at grade 4 were assigned holistically (a single score per event). At grades 8 and 12, each performance event score was the sum of multiple scored items within the task. Within grades 8 and 12, the number of total points that could be earned on a performance event varied substantially.

CONSISTENCY OF SCORING. Scoring of performance events was monitored through a quality assurance process similar to that used for open-response questions. Each training leader reviewed at least two responses per scorer each morning and afternoon as a way of monitoring scorer consistency. However, performance events were not routinely subjected to the two percent rescoring used for open-response items.

As an independent check on performance event scoring, a separate rescoring study was conducted on grade 8 performance events from three areas: mathematics, science, and arts & humanities. Following completion of regular scoring, two hundred papers per content area were scored twice. Across all six hundred papers, scores assigned by two different scorers correlated .91, with 86% of scores falling within one point of each other, 95% falling within two points, and more than 99% falling within three points. Maximum obtained scores ranged from 6 points for one of the arts & humanities performance events to 12 points for science and mathematics performance events. Interestingly, but not surprisingly given their relatively short range of scores, arts & humanities performance events yielded a lower correlation across scorers (.66) than did science (.87) and mathematics (.98).

MATHEMATICS, WRITING AND ALTERNATE PORTFOLIOS

Teacher training, scoring and auditing are described for portfolios in Chapter 12.

CHAPTER 9
EQUATING AND LINKING

INTRODUCTION

To obtain meaningful results across multiple years of the KIRIS Assessment, it is necessary that the scores from one year mean the same thing (i.e., can be interpreted the same way) as scores from previous years. In an accountability system that depends upon comparing growth from one year to the next, measurement of such growth would not be possible if the scores from previous years had a different meaning than the scores for the current year. Attempting to measure growth under such circumstances would be both inaccurate, and more importantly, misleading. Clearly, for decisions regarding school level accountability, the scores obtained across multiple years of an assessment must mean the same thing or in psychometric terms, must be equated, placed upon the same scale of measurement.

The implementation of KIRIS requires several types of linking, due to changes in the test items for each particular year, matrix-sampling of items in each year, and changes in the scaling model and grades tested over the course of the accountability cycle.

- Each year, 12 forms of the KIRIS assessment are administered to students at each accountability grade. Because each of the 12 forms contains two matrix sampled items per content area unique to a particular form, the 12 different forms are not identical. As such, a scaling model based in Item Response Theory (IRT) was used to place scores from the different forms collectively onto the same scale.
- Another type of linking implemented in KIRIS involves taking the 12 forms from one year of the assessment and placing them on the same scale as the 12 forms from the previous year's assessment. Because of this equating, meaningful decisions regarding growth across years can be made. Part of this equating involves determining whether it is necessary to make adjustments for changes in scorer leniency from one year of the assessment to the next, and, if changes have occurred, to adjust accordingly. It is highly preferable for scoring standards to remain constant across accountability years, particularly for items that are administered in two different years and used to equate scores across the years, but sometimes necessary to make adjustments.
- A third type of linking involved placing the proficiency level cut scores from one Accountability Cycle I onto the scale used in Accountability Cycle 2. This linkage was necessitated by the scaling changes made in Cycle 2 (i.e., common and matrix items were scaled together using IRT so as to produce one score). In order to implement a standards validation process, the relationship between the original and revised scales needed to be assessed using a common person equating.
- A fourth type of linking involves taking into account differences in grades tested, in that KIRIS testing of high schools shifted from testing of grade 12 students exclusively from 1991-1992 to 1993-94, to testing grade 11 students on open-response and grade 12 students on portfolios for 1994-95 and 1995-96. This was accomplished at the school level, by giving schools a slightly lower baseline to reflect the finding that grade 11 students scored lower than 12th graders.

The present chapter discusses the procedures required for implementing the latter three types of linking mentioned above. The first type of linking, scaling responses together using common items to link matrix-sampled forms, is accomplished through a relatively straightforward application of Item Response Theory, and is not discussed in this chapter. Instead, the first section describes the results of linking four years of open-response tests in the content areas, reading, mathematics, science and social studies. The document, "Equating Plan: Accountability Cycle 2" (Appendix 9-1) describes in detail the procedures used to obtain these results. These procedures reflect a slight modification to the equating procedures outlined in an earlier version of the document, *Equating Plan: Accountability Cycle 2* (Kingston, 7/5/95), based on recommendations of the KIRIS National Technical Working Group (NTWG).

In the second section of this chapter the procedures used to link the cut scores between Accountability Cycle I and Accountability Cycle 2 are presented. The need to link proficiency levels to Accountability Cycle I arose when the Department of Education, upon recommendation of the NTWG, decided to undertake a standard validation process rather than to set new standards for Accountability Cycle 2.

The third section describes results of the process used to link open-response items in Arts & Humanities (AH) and Practical Living/Vocational Studies (PL) across the years of Accountability Cycle 2. Linking tests in these areas differs from equating the open-response portions of the assessments in reading, mathematics, science, and social studies due to several differences in the design of the AH and PL tests relative to the other content areas.

The fourth section contains analyses designed to estimate the standard error of equating associated with each grade level and content area in KIRIS. With respect to the equating procedures outlined in this chapter, it is important to determine the level or magnitude of equating error that may be associated with the procedures. This section describes a study designed to assess errors in equating via jackknife procedures.

The final sections of this chapter describe procedures used to adjust baseline scores to account for moving open-response testing from grade 12 to 11, and results of efforts to equate performance event scores. Note that there are several appendices related to this chapter.

OPEN RESPONSE TESTS IN READING, MATHEMATICS, SCIENCE, AND SOCIAL STUDIES

INTRODUCTION

This section provides a summary of the results of linking the open-response portions of the KIRIS assessments in Accountability Cycle 2 (years 1992-93, 1993-94, 1994-95, and 1995-96). Each content area/grade/year combination (e.g., Reading, Grade 8, 1993-94) will be referred to as a test. There are twelve tests per year, each of which needed to be equated to the base year, 1992-93, for a total of 36 equatings (four content areas, three grades, three equatings apiece). For the sake of brevity, the school years 1992-93, 1993-94, 1994-95, and 1995-96 will be referred to as 93, 94, 95, and 96, respectively.

The tests were linked using a revised version of the procedures described in *Equating Plan: Accountability Cycle 2* (Kingston, 7/5/95), for which a simplified rule to eliminate outliers was applied. Additionally, equating constants were applied to student ability estimates rather than item parameters. All item calibrations and initial student abilities were estimated using MULTILOG® (1991). Samejima's (1969) two-parameter graded response model for ordered responses was used to estimate item parameters. Additional analyses were conducted using the SAS® software system.

ITEM CALIBRATION SAMPLES

A calibration sample was selected for each year/grade combination. Each sample included more than three-fourths of the entire population of students receiving scores on the assessment. Because of the large sample sizes, it was not necessary to create a separate calibration sample for each test within a year/grade. Students were excluded from the calibration sample for the following reasons:

- Student excluded from the assessment for approved reason,
- Student did not respond to one or more items on a test in any of the four content areas,
- Student obtained a total raw score of 0 on a test in any of the four content areas,
- Student record did not have a valid form number (i.e., 01-12; generally non-tested students who did not participate in the testing but were not excluded),
- Student did not take the same form of the test in all four content areas.

Table 9-1 provides the number of students included in and excluded from each calibration sample. The reason for exclusion is also provided. Students who would have been excluded for more than one of the five reasons listed above are assigned to the exclusion category which occurs first in that list. For example, a student who did not respond to 3 items on the mathematics test and received a total raw score of 0 on the science test would be counted as excluded from the calibration sample for not responding to one or more items.

There was one problem in applying these exclusion rules that could impact the results of the equating. As revealed by the data in Table 9-1, in the 93 sample it was not possible to exclude students for non-response to an item. This is due to the scoring rules that were in effect through 93. Previously, blank responses were assigned the lowest score point (0). Removing all of these cases from the calibration sample would leave only 4 score points (1, 2, 3, and 4). Beginning in 94, blank responses were assigned a score of B and totally incorrect responses were assigned the lowest score point (0). Removing the B scores from the calibration sample still left cases of all score points in the calibration sample. For the purpose of computing ability estimates after calibration, scores of B were assigned a value of 0.

SAMPLE CHARACTERISTICS. Each calibration sample contains students who completed 1 of 12 test forms. In each content area, each test form contained a set of items common to all forms and a set of matrix-sampled items unique to the form. There were five common items at all grades in 93 and 94 and for grade 4 in 95 and 96. There were six common items at grades 8 and 11 in 95 and 96. In all years, there were two matrix sampled items administered per content area.

Because the administration of test forms was spiraled within classroom, there should be an approximately equal number of students completing each test form and the ability of students completing each test form should be equivalent. For each calibration sample, the number of students completing each test form and the common item raw score distributions by test form were recorded.

ITEM CALIBRATION RESULTS

Item parameters were estimated with Samejima's (1969) two-parameter graded response model for ordered responses using MULTILOG®. For each five-category open-response item a single discrimination parameter and four difficulty parameters are estimated.

MULTILOG® software allows a maximum of 25 cycles to estimate the item parameters under marginal maximum likelihood estimation. The default cycle convergence criterion is 0.001. For each test, the number of cycles required to estimate the item parameters and the maximum intercycle parameter change after the final cycle are presented in Table 9-2. Grade 12 Science in 93 is the only test that failed to approach the convergence criterion. This appears to be due to several items on which very few students achieved a score of 4.

RESCORING ANALYSIS

As part of the linking process, a rescoring analysis was conducted to evaluate the extent of drift in the scoring standards from year to year. As described in the equating plan, a sample of papers for reused items from the previous year are rescored during the operational scoring session in the current year. For each item, recast means and standard deviations of the two sets of scores were compared. Also, a weighted correlation between the original and rescores was computed.

After examining the results of the rescoring analysis, it was determined for 94 and 95 that any minor changes in scoring that may have occurred on a single item were balanced across items within a content area. There did not appear to be a consistent shift in scoring in either a positive or negative direction within a content area. Based on these results, no adjustments to scores were needed. (It should be noted that because of the scoring change instituted in 94 , 93 papers that were rescored either 0 or 1 in 94 were classified as receiving a rescore of 1 for comparisons with original 93 scores.) For 96, however, it was necessary to make adjustments in some content areas. The procedures used to obtain the adjustments, given in Table 9-3, are described in Appendix 9-2.

TABLE 9-1 CALIBRATION SAMPLES									
Grade-Year	Total Population	Included in Sample	Excluded: Valid Exc	Excluded: Non Resp	Excluded: 0 Score	Excluded: Inv Form	Excluded: Mult Form	% of Total Population in Sample	% of Students Rec Scores in Sample
4 - 93	52242	48705	2221	174	1030	45	67	93.2	97.3
4 - 94	50296	42947	510	5954	681	0	204	85.4	86.3
4 - 95	50321	46239	864	2610	573	1	34	91.9	93.5
4 - 96	48482	44986	723	2201	541	0	31	92.8	94.2
8 - 93	53045	47689	2195	248	2698	59	156	89.9	93.8
8 - 94	52596	42108	655	8244	1546	0	43	80.1	81.1
8 - 95	51974	46193	755	4439	496	0	91	88.9	90.1
8 - 96	51640	45402	939	4435	831	0	33	87.9	89.5
12 - 93	40191	34051	1690	454	3364	152	30	84.7	88.4
12 - 94	37390	28726	629	7419	605	0	11	76.8	78.1
11 - 95	42095	34892	398	6231	556	0	18	82.9	83.7
11 - 96	41924	34266	546	6641	462	0	9	81.7	82.8

Note: % of Total Population in sample is computed by dividing (Included in Sample) by (Total Population);
Students receiving scores are all students included in the computation of the accountability index;
% of Students receiving scores in sample is computed by dividing (Included in Sample) by (Total Population - Excluded: Valid Exc).

TABLE 9-2
NUMBER OF CYCLES AND MAXIMUM PARAMETER CHANGE OF FINAL CYCLE OF ITEM PARAMETER ESTIMATION, BY YEAR

Grade and Content Area	93	94	95	96
Grade 4 Reading	23 (0.00120)	20 (0.00098)	18 (0.00092)	14 (0.00099)
Grade 4 Mathematics	23 (0.00065)	21 (0.00080)	20 (0.00080)	22 (0.00077)
Grade 4 - Science	19 (0.00050)	19 (0.00062)	19 (0.00064)	21 (0.00059)
Grade 4 - Social Studies	17 (0.00074)	17 (0.00056)	19 (0.00051)	20 (0.00067)
Grade 8 - Reading	24 (0.00098)	22 (0.00074)	21 (0.00027)	23 (0.00070)
Grade 8 - Mathematics	21 (0.00114)	25 (0.00087)	25 (0.00152)	25 (0.00165)
Grade 8 - Science	24 (0.00068)	21 (0.00105)	25 (0.00069)	25 (0.00073)
Grade 8 - Social Studies	25 (0.00232)	22 (0.00054)	25 (0.00193)	25 (0.00185)
Grade 12 - Reading	24 (0.00059)	21 (0.00060)	22 (0.00103)	25 (0.00076)
Grade 12 - Mathematics	21 (0.00095)	25 (0.00167)	25 (0.00263)	25 (0.00168)
Grade 12 - Science	25 (0.09694)	24 (0.00075)	25 (0.00132)	25 (0.00075)
Grade 12 - Social Studies	25 (0.00134)	24 (0.00096)	25 (0.00225)	25 (0.00191)

Notes: The information provided is # cycles (max change); Grade 11 replaces grade 12 in 95 and 96.

TABLE 9-3
ADJUSTMENT FOR SCORER DIFFERENCES, AS APPLIED TO 1995-96 SCORES

CONTENT AREA	GRADE 4	GRADE 8	GRADE 11
Reading	NA	NA	NA
Mathematics	NA	NA	NA
Science	.1209	NA	NA
Social Studies	.1304	.0237	.2859
Arts & Humanities	-.1517*	.0852	-.3070
Practical/Vocational	NA	-.1127	-.4632

* Negative differences indicate that scoring was more lenient in 1995-96 than in 1994-95.

EQUATING RESULTS

After the initial calibration of item parameters, equating procedures were conducted to find the linear transformation necessary to place all item parameters and ability estimates on the appropriate 93 metric. As described in the equating plan, this was accomplished by first linking the 94 test to the 93 test to place the 94 parameters on the 93 metric. Then, using the 94 adjusted parameters, the 95 test was linked to the 94 test and consequently to the 93 test. Lastly, using the 95 adjusted parameters, the 96 test was linked to the 95 test and consequently to the 93 test. As such, the end result of the above procedures was to place all final adjusted parameters onto the same 93 metric. The locations of linking items that served as the basis of the equating are presented in Appendix 9-3, except the linkages for the first year of the accountability cycle, which are documented in the KIRIS Accountability Cycle I Technical Manual.

Before presenting the summary of equating results based upon the procedures outlined in "Equating Plan: Accountability Cycle 2" (Appendix 9-1), there are three points concerning the equating procedures which should be noted:

1. When reused item score points are first removed from the equating in Step 8E of the equating plan, the procedures call for the entire item to be removed unless the total number of b's is less than 40, in which case only the affected pair of b's (i.e., both the predecessor year b and the current year b) should be removed. In every case, removing entire items would have resulted in fewer than 40 b's remaining. Therefore, in all cases, only the affected pair of b's were removed.

2. Because of scoring changes implemented in 94, it was not possible to use the lowest difficulty from each item (0 vs. 1, 2, 3, 4) to link the 94 and 93 tests. Only the second, third, and fourth b's could be used in the equating. This was not the case in linking the remaining tests.

3. Common items that had not been edited were included in the equating when the number of remaining b's was less than 32 after adjustments of matrix item parameters and deletions of outliers. The grade 4 science test was handled differently because there were only 9 items used in both years 93 and 94, including common items in 94. Only 27 b's were available from the beginning of the process and all were used in the equating. For grade 8 science and grade 4 social studies in 96, only eight items (32 score points) were initially available for equating. In this case, a .6 cutoff criterion was used (as opposed to .4, see Step 8E), which was consistent with that of Arts and Humanities and Practical Living/Vocational Studies.

Equating results are summarized in Table 9-4. The final slopes and intercepts presented yield the transformations necessary to place item parameters and ability estimates from 94, 95, and 96 onto the 93 scale. Final slopes and intercepts were applied to the student ability estimates for all students receiving scores, not just the calibration sample. Distributions of scores in this ability estimate metric, which provide a somewhat less complicated scale for assessing growth than the standards-based accountability scores and are sometimes referred to as theta, across the four years of the second accountability cycle are presented in Table 9-5.

TABLE 9-4
SUMMARY OF EQUATING RESULTS

Grade	Content	Years	# reused matrix items	initial # b's	Common items used (y/n)	# b's remaining	Final slope	Final intercept	Correlation between reused b's	Mean (sd) reused b's
4	Rdg	93/94	10	30	y	38	0.842	0.388	0.987	1.310 (1.299)
	Mth	93/94	12	36	y	41	0.832	0.473	0.985	1.286 (1.045)
	Sci	93/94	9	27	y	27	0.877	0.250	0.939	1.302 (1.521)
	Soc	93/94	13	39	n	34	0.981	0.233	0.989	1.280 (1.438)
4	Rdg	94/95	9	36	y	32	0.897	0.842	0.990	0.939 (1.661)
	Mth	94/95	10	40	n	32	0.770	0.640	0.993	0.976 (1.121)
	Sci	94/95	10	40	y	39	0.829	0.629	0.995	0.916 (2.009)
	Soc	94/95	9	36	y	38	1.067	0.363	0.995	0.472 (2.175)
4	Rdg	95/96	11	44	n	41	0.869	0.872	0.996	0.856 (1.764)
	Mth	95/96	12	48	n	47	0.748	0.595	0.990	0.872 (1.120)
	Sci	95/96	12	48	n	40	0.797	0.534	0.995	1.130 (2.017)
	Soc	95/96	8	32	n	28	0.874	0.313	0.995	0.481 (2.436)
8	Rdg	93/94	13	39	n	38	0.951	0.460	0.989	1.143 (1.234)
	Mth	93/94	17	51	n	38	0.839	0.370	0.982	1.513 (1.052)
	Sci	93/94	16	48	n	40	0.822	0.144	0.991	1.407 (1.322)
	Soc	93/94	17	51	n	40	0.964	0.396	0.989	1.749 (1.208)

TABLE 9-4, CONTINUED

Grade	Content	Years	# reused matrix items	initial # b's	Common items used (y/n)	# b's remaining	Final slope	Final intercept	Correlation between reused b's	Mean (sd) reused b's
8	Rdg	94/95	10	40	y	46	0.788	0.386	0.993	0.425 (1.761)
	Mth	94/95	10	40	n	37	0.819	0.537	0.994	1.279 (1.113)
	Sci	94/95	11	44	n	32	0.811	0.206	0.994	0.701 (1.617)
	Soc	94/95	9	36	n	32	1.039	0.459	0.996	1.248 (1.883)
8	Rdg	95/96	11	44	n	39	0.721	0.416	0.995	0.566 (1.764)
	Mth	95/96	12	48	n	46	0.751	0.533	0.993	1.204 (1.265)
	Sci	95/96	8	32	n	30	0.715	0.003	0.992	1.056 (1.528)
	Soc	95/96	11	44	n	36	0.881	0.266	0.994	0.817 (1.761)
12	Rdg	93/94	11	42	n	35	0.981	0.608	0.989	1.097 (1.174)
	Mth	93/94	17	51	n	35	0.871	0.539	0.991	1.666 (0.985)
	Sci	93/94	19	57	n	44	0.954	0.479	0.990	2.089 (1.512)
	Soc	93/94	17	51	n	39	0.949	0.888	0.989	1.656 (1.361)
12/11	Rdg	94/95	8	32	y	35	0.930	0.559	0.992	0.852 (1.691)
	Mth	94/95	11	44	n	41	0.923	0.381	0.983	1.292 (1.414)
	Sci	94/95	11	44	y	40	0.924	0.602	0.993	1.241 (1.666)
	Soc	94/95	9	36	y	43	1.030	0.607	0.994	0.797 (1.825)
11	Rdg	95/96	10	40	y	33	0.874	0.559	0.990	1.004 (1.436)
	Mth	95/96	9	36	n	34	0.832	0.519	0.992	1.272 (1.319)
	Sci	95/96	9	36	n	32	0.870	0.608	0.992	1.327 (1.577)
	Soc	95/96	12	48	n	32	0.969	0.440	0.990	1.460 (1.601)

Note: Where applicable, scorer adjustments noted in a preceding section are included in the Final Intercept figures.

TABLE 9-5
DISTRIBUTION OF STUDENT ABILITIES IN ACCOUNTABILITY CYCLE 2

Grade, Content Area	Mean (Standard Deviation), By Year			
	93	94	95	96
Grade 4 - Reading	-0.027 (0.952)	0.272 (0.823)	0.777 (0.865)	0.767 (0.901)
Grade 4 - Mathematics	-0.020 (0.935)	0.390 (0.779)	0.586 (0.727)	0.523 (0.728)
Grade 4 - Science	-0.029 (0.918)	0.161 (0.837)	0.569 (0.769)	0.447 (0.781)
Grade 4 - Social Studies	-0.031 (0.937)	0.107 (0.952)	0.283 (1.019)	0.210 (0.895)
Grade 8 - Reading	-0.063 (1.016)	0.242 (1.028)	0.290 (0.820)	0.270 (0.831)
Grade 8 - Mathematics	-0.055 (0.994)	0.221 (0.789)	0.458 (0.790)	0.420 (0.756)
Grade 8 - Science	-0.062 (0.992)	0.001 (0.763)	0.122 (0.786)	-0.108 (0.712)
Grade 8 - Social Studies	-0.059 (1.010)	0.168 (1.029)	0.325 (1.063)	0.112 (0.920)
Grade 12 - Reading	-0.087 (1.016)	0.401 (1.015)	0.416 (0.950)	0.372 (0.960)
Grade 12 - Mathematics	-0.090 (1.004)	0.382 (0.835)	0.417 (0.845)	0.367 (0.837)
Grade 12 - Science	-0.107 (1.045)	0.276 (0.952)	0.464 (0.922)	0.433 (0.886)
Grade 12 - Social Studies	-0.127 (1.069)	0.457 (1.031)	0.442 (1.037)	0.246 (0.995)

Note: Grade 11 replaces Grade 12 in 95 and 96.

LINKING PROFICIENCY LEVEL CUTPOINTS TO ACCOUNTABILITY CYCLE I

Given the decision to validate Accountability Cycle I standards rather than reset them entirely for Cycle 2, there was a need to link the proficiency level cutpoints across cycles, across the additive scoring model/IRT hybrid used in Cycle I and the IRT model implemented for Cycle 2. Two options for accomplishing this were considered:

1. Re-equate the assessments using a common items, non-equivalent groups design, with 1991-92 as the base year.
2. Determine the relationship between the original and revised 1992-93 scales using a common person method, and adjust the proficiency level cutpoints accordingly.

Option 1 was rejected for a variety of reasons. Among the reasons were the well-documented concerns related to the length of the tests in 1991-92, the small number of reused items for equating, and consequently the need to use both common and matrix items for equating. The 1991-92 and 1992-93 assessments were equated once using a common items, non-equivalent groups design out of necessity. If not necessary, it was considered desirable to avoid equating the assessments in this manner again.

A plan for Option 2 was developed using common person equating. The process involved a linear transformation of cutpoints from the original scale to the revised scale.

1. Student scores were computed using the new parameters and the mean and standard deviation were computed.

2. Student scores were computed using the original parameters and the mean and standard deviation were computed.

3. A linear transformation was applied to convert the original cutpoints to the new scale using the formula: $ADJ = [(ORIG - MN_{(ORIG)})/SD_{(ORIG)}] * SD_{(NEW)} + MN_{(NEW)}$

For example, the original Grade 4 Reading proficiency level cutpoints were as follows:

N/A:	-0.10
A/P:	1.46
P/D:	2.60

The mean and standard deviation of student scores on the original scale were 0.301 (0.861). The mean and standard deviation on the revised scale were -0.027 (0.952). The adjusted P/D cutpoint was computed to be 2.52. Adjusted cutpoints for all content areas are presented in Appendix 9-4.

LINKING ARTS & HUMANITIES AND PRACTICAL LIVING/VOCATIONAL STUDIES ACROSS YEARS

A modified version of the procedure used to equate reading, mathematics, science, and social studies was developed to link Arts & Humanities (AH) and Practical Living/Vocational Studies (PL) across and within years, due to the following ways in which the AH and PL tests differ in their design from the other content areas:

- During all years of the assessment the items were totally matrix sampled;
- In 1992-93 and 1993-94, most students completed either one or two items per content area and some students did not complete any items within a content area;
- In 1992-93 and 1993-94, items were embedded within reading, mathematics, science, and social studies portions of the assessment; and,
- In 1992-93 and 1993-94, most items received a single score that was applied to both the AH or PL test and the content area in which the item was embedded.

Beginning in 1994-95, a separate test section was created for AH and PL. All students completed two items in each content area, and items were totally matrix sampled and scored separately from reading, mathematics, science and social studies.

Linking the Arts & Humanities (AH) and Practical Living/Vocational Studies (PL) tests across years was accomplished through a modification of the common item equating procedures used to equate the open-response tests. The two modifications made to the original process were in the procedure for estimating the initial difficulties and the cutoff for eliminating items from the equating step.

Initial item difficulties were estimated by computing a log-odds ratio based on the cumulative percentage of students scoring each point or above. For each item four difficulties were computed corresponding to the percentage of students scoring 1 or above, 2 or above, 3 or above, or 4. This process has been used to estimate item difficulties for AH and PL items since the initial year of testing in those content areas (i.e., 1992-93). An example is provided in Table 9-6.

TABLE 9-6
EXAMPLE OF COMPUTING THE INITIAL ESTIMATES OF AH AND PL ITEM DIFFICULTIES

Scores	% scoring (P)	% not scoring (Q)	Difficulty: ln(Q/P)
1, 2, 3, or 4	90.6	09.4	-2.27
2, 3, or 4	47.0	53.0	0.12
3 or 4	14.0	86.0	1.82
4	02.8	97.2	3.55

Based on the differences in item difficulties from year to year, the cutoff for eliminating items from equating was raised from 0.40 to 0.60 because of the limited number of items available across years. The lack of items available for linking the tests was due in large part to the length of the tests. The AH and PL tests are much shorter than the other open response tests. In 93 and 94, the AH tests contained 9 items (with the exception of grade 8 in 93 which had 6 items available) and the PL tests had 11 items. In 95 each test contained 12 items and in 96 each test contained 24 items. The number of reused items and available linking points are provided in the Table 9-7.

In all cases prior to 96, the starting number of linking points was smaller than (or in one case, equal to) the minimum number of 32 points used in linking the open-response tests in reading, mathematics, science and social studies. Using the 0.40 criterion, the remaining number of linking items would have dropped below 16 in several cases. The number of remaining linking items using the 0.60 and 0.40 cutoffs are presented in the Table 9-8. Table 9-9 provides equated mean and standard deviations of the difficulties of the remaining linking items.

TABLE 9-7
NUMBER OF ITEMS AVAILABLE FOR LINKING TESTS

Grade, Content Area	94 to 93		95 to 94		96 to 95	
	# reused items	# linking points	# reused items	#linking points	# reused items	# linking points
Grade 4 AH	7	21	4	16	6	24
Grade 4 PL	9	27	6	24	5	20
Grade 8 AH	6	18	6	24	6	24
Grade 8 PL	9	27	6	24	6	24
Grade 12 AH	7	21	6	24	7	28
Grade 12 PL	7	21	6	24	8	32

TABLE 9-8
NUMBER OF LINKING POINTS AVAILABLE WITH .60 AND .40 CUTOFF CRITERION

Grade, Content Area	94 to 93		95 to 94		96 to 95	
	.60	.40	.60	.40	.60	.40
Grade 4 AH	20	18	15	12	20	18
Grade 4 PL	19	13	17	12	19	14
Grade 8 AH	11	8	15	11	23	18
Grade 8 PL	27	27	12	5	18	13
Grade 12 AH	18	13	16	14	21	19
Grade 12 PL	15	13	15	11	20	11

TABLE 9-9
MEAN AND STANDARD DEVIATION OF REMAINING LINKING ITEMS

Grade, Content Area	94 to 93	95 to 94	96 to 95
Grade 4 AH	1.753 (1.690)	0.953 (2.271)	1.644 (2.527)
Grade 4 PL	2.212 (1.774)	1.203 (2.478)	1.417 (2.701)
Grade 8 AH	1.638 (1.983)	0.297 (2.370)	0.933 (2.348)
Grade 8 PL	2.466 (1.990)	-0.038 (1.478)	0.982 (2.126)
Grade 12 AH	2.459 (2.003)	1.100 (2.405)	0.509 (2.701)
Grade 12 PL	2.364 (1.791)	1.493 (2.600)	0.250 (3.110)

Note: Grade 11 replaced Grade 12 in 95 and 96

ERROR OF EQUATING

With respect to the equating procedures discussed in this chapter, it is important to determine the level or magnitude of equating error associated with such procedures. This section describes a study designed to assess errors in equating via jackknife procedures (Mosteller and Tukey, 1977). More specifically, analyses were employed to estimate the standard error of equating associated with each grade level and content area combination.

To estimate the standard error of equating, the following "*n*" steps were used to obtain multiple sets of equating constants for each grade and content area:

1. Drop the first linking item and recalculate equating constants.
2. Add the first linking item back in. Drop the second linking item and recalculate equating constants.
3. Add the second linking item back in. Drop the third linking item and recalculate equating constants.

.

.

.

n. Add the (n-1) linking item back in, drop the n^{th} linking item and recalculate equating constants. After *n* equating constants have been calculated, apply the constants (linear transformation using means and standard deviation equating) to different values of raw or unequated theta (i.e., -3.0, -2.5, -2.0,..., 2.5, 3.0).

The preceding steps provide multiple estimates of the equating constants which were subsequently applied at each value of raw theta noted above. While the final constants used for equating scales across years are based upon all linking items from the original set of items, each of the *n* sets differs by one linking item. As such, the variability across the different estimates reflects the error one might expect had a somewhat different set of linking items been used to calculate final equating constants. Because the magnitude of equating error typically varies across the theta scale for a linear equating method, the variability at different levels of theta was obtained to assess the pattern of equating error from low to high values of theta. To obtain estimates of this variability at each level of theta, the following jackknife procedure was applied:

Let y_{all} represent the original equating constants applied at a given level of theta, and y_i represent the aforementioned *n* equating constants (with item i omitted) applied at a given level of theta, then the i^{th} pseudovalue, denoted p_i is given by:

$p_i = n\, y_{all} - (n - 1)\, y_i$ where n is the number of cases (or items).

While the jackknifed estimate is the average of the n pseudovalues at a particular level of theta, the standard error (or the measure of variability of interest) is obtained by first calculating the variance of the n pseudovalues, and then dividing the variance by n and taking the square root. The process was repeated for each level of raw or unequated theta (i.e., -3.0, -2.5, -2.0,..., 2.5, 3.0). More specifically, p, was obtained at different levels of theta to assess the pattern of equating error from low to high values of theta.

Figures 9-1, 9-2 and 9-3 present estimates for the standard error of equating for each grade and content area across the theta scale from -3 to +3. Several important observations can be made with regard to the patterns of equating error presented in Figures 9-1 to 9-3. First, the overall magnitude of the standard error of equating estimates appears to vary across the theta scale, with lower values generally in the middle of the scale and higher values toward the ends of the scale. These results are consistent with the expectations of a linear (i.e., means and standard deviation) equating method, i.e., the least amount of variability occurs at the mean, with equating error larger at the tails. Secondly, in general, the overall magnitude of the standard error of equating estimates is small to moderate in relation to theta standard error estimates which typically have values from .40 to .50. The magnitude of the standard error of equating for each grade level and content area generally ranged from .05 to .10 toward the middle of the theta scale and from .05 to .20 toward the ends of the theta scale. Thirdly, the standard error of equating appears to vary by content area. That is, the standard error of equating for reading and mathematics tended to be smaller in magnitude than for science and social studies. Finally, the impact of the number of items upon which equating constants were based seem to be secondary in importance relative to the differences observed among the content areas. For example, although the number of initial items appears to be an important factor relative to the higher standard errors for grade 4 social studies (8 items), higher values were also observed in grade 11 social studies where the number of initial items far exceeds that for the other content areas (13 items).

FIGURE 9—1

EQUATING ERROR ACROSS JACKKNIFE ESTIMATES, GRADE 4

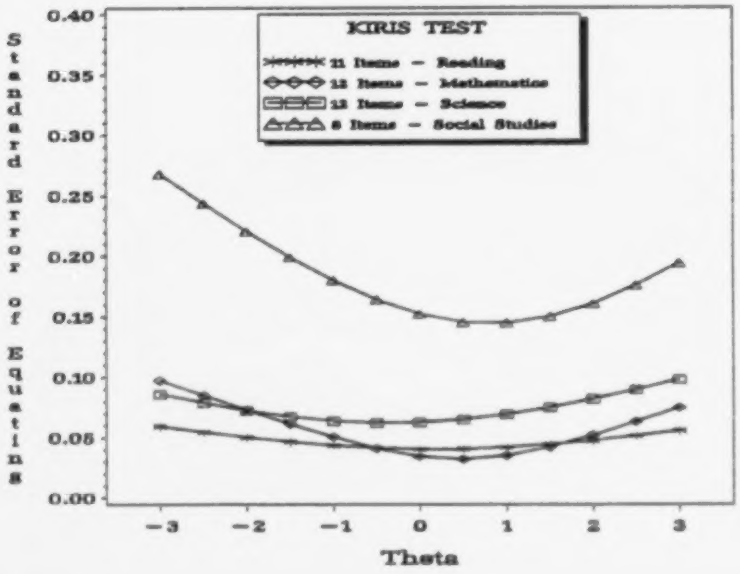


FIGURE 9–2
EQUATING ERROR ACROSS JACKKNIFE ESTIMATES, GRADE 8

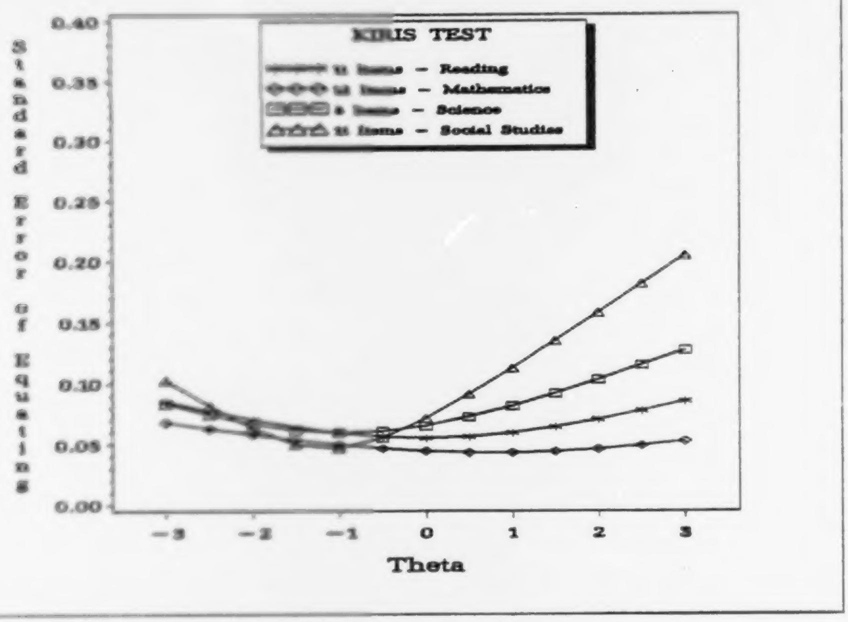
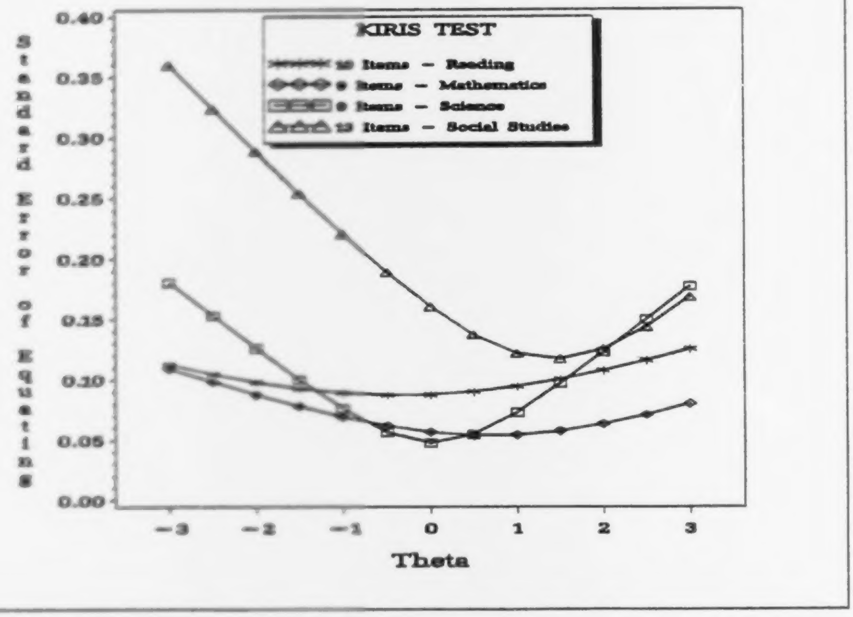


FIGURE 9–3
EQUATING ERROR ACROSS JACKKNIFE ESTIMATES, GRADE 11



BASELINE ADJUSTMENT TO ACCOUNT FOR ADMINISTRATION OF OPEN RESPONSE TO GRADE 11 STUDENTS RATHER THAN GRADE 12 STUDENTS

Due to administration of open-response tests in grade 11 in 1994-95 and 1995-96 as compared with grade 12 in 1992-93 and 1993-94, it was necessary to adjust 1992-93 and 1993-94 scores to reflect differences in the achievement of eleventh and twelfth grade students. To provide an accurate estimate of grade 12 student scores, all twelfth graders took the tests and a sample of responses from one-sixth of the twelfth graders taking the KIRIS assessment in 1994-95 was drawn at random and scored with the grade 11 student responses in Reading, Mathematics, Science, Social Studies, Arts and Humanities, and Practical Living/Vocational Studies. After eliminating scores for students excluded from school accountability index calculations, 6238 twelfth grade students remained. For these students' scores, Novice, Apprentice, Proficient and Distinguished classifications were converted to the zero to 140 scale used in calculating school accountability indices, and compared against scores for grade 11 students. As shown in Table 9-10, twelfth graders outperformed eleventh graders in every case, by as little as one-tenth of a point in Science to as much as three points in Reading.

To afford an appropriate comparison with grade 11 scores for 1994-95, grade 12 scores shown on statewide student performance reports for 1992-93 and 1993-94 were reduced by the point difference between the grade 12 sample and the grade 11 results for 1994-95. For accountability purposes, the pertinent reduction to baseline is the number of points difference multiplied by the weight of the open-response components in calculating the school accountability index. For reading, social studies and science, scores on open-response questions are each given a weight of .14. Half this weight, or .07 is assigned to each of the two content areas, Arts & Humanities and Practical Living/Vocational Studies. Finally, for Mathematics, .14 total weight is assigned, but only seven-tenths of this comes from open-response questions, giving open-response questions a total weight of .098 in Mathematics. Multiplying these weights by the differences in each content area yields a total of .7, when rounded to the nearest tenth. (Note that weights do not sum to unity; the remainder is covered by portfolios and non-cognitive indicators, which were not changed with regard to the grade upon which high school accountability baselines were computed.) The sum of the weighted differences, rounded to the nearest tenth, .7, was the amount by which every high school's baseline was adjusted.

PERFORMANCE EVENTS

Following criticism of performance event equating undertaken in the first accountability cycle, KDE and ASME jointly examined alternatives to the approach applied to place 1993-94 and 1992-93 performance events on the original 1991-92 scale. The approaches initially considered relied upon open-response question scores as an external anchor. One possibility was to match individual student performance event scores (in nearly all cases, each student took one performance event) with open-response scores, and to link performance event scores across years through the open-response test equating. After matching student scores in this regard (a time-consuming process requiring, as a first step, hand entry of student names onto

TABLE 9-10
GRADE 11 POPULATION AND GRADE 12 SAMPLE 1994-95 CONTENT AREA ACCOUNTABILITY INDEX SCORE MEANS AND DIFFERENCES

Content Area	Grade		Mean Difference		Weight in accountability index
	11 (N-41469)	12 (n-6238)	Raw	Weighted	
Reading	33.62	36.68	3.03	.14	.424
Mathematics	45.40	47.41	2.00	.098	.196
Science	40.75	40.86	0.11	.14	.015
Social Studies	39.92	40.08	0.16	.14	.026
Arts & Humanities	16.99	17.31	0.32	.07	.022
Practical Living/ Vocational Studies	25.14	25.98	0.84	.07	.059

a computer data file), it was discovered that the correlation between performance event and open response scores was too low to support use of an external anchor at the student level.

A variation on using open response as an external anchor at the student level is to make the strong assumption that statewide changes performance event and open response scores in a given content area should mirror each other. To the extent that open response questions and performance events are measuring highly related constructs, and scored on the basis of written answers to multi-part questions, this assumption is not entirely unreasonable. Implementing the external anchor approach at the state level, however, would have the undesirable effect of ignoring any differential gains in these two areas. If open-response scores in science, for example, were unchanged from year, then science performance event scores would be assumed to be unchanged as well. Thus, even if all schools had made great strides in enhances student skills necessary to succeed on science performance events, about half the schools' scores would decline and half would rise. This approach was dismissed after initial consideration, as was replication of the hybrid common-person, randomly equivalent groups design followed in Cycle I. The latter was dropped for want of obtaining appropriately large samples necessary for the equating to succeed as well as some question as to the error that would be introduced by using out of state students.

Having effectively exhausted the data-based equating methods available for linking performance event scores, a judgment-based method was developed after consultation with NTWG. Despite every effort to implement the method selected, draft results proved untenable, yielding unbelievable drops in some content areas and equally unbelievable gains in others. As a result, 1994-95 performance events scores were never finalized, and were not reported to schools by KDE or ASME. Given the lack of a satisfactory equating, performance events for the final year of the accountability cycle were not scored. For a description of the judgment-based methods and the initial results obtained, see Appendix 9-5.

CHAPTER 10
REPORTING TO SCHOOLS AND DISTRICTS

The Kentucky Department of Education (KDE) notifies schools and districts of their KIRIS performance results during the second month of each school year. This distribution includes detailed descriptions of student level and content area scores that lead to a school's and district's performance judgment. At the end of a four year cycle, KDE notifies each school and district of their single performance judgment.

Final results for the second accountability cycle , a four year period beginning with the 1992-93 school year and ending at the conclusion of the 1995-96 school year, were released to schools and districts on October 18, 1996. Since the second accountability cycle of KIRIS testing was completed, schools' and districts' overall accountability scores were reported. These scores were compared to each school's/district's improvement goal to determine their level of success. Each school's or district's accountability score for Cycle 2 was translated into a school's or district's performance judgment.

A school or district whose final accountability growth index, which is the weighted average of the 1994-95 and the 1995-96 accountability indices, exceeded its Cycle 2 improvement goal by one point or more and had met the requirement of reducing its percentage of Novices were judged in a reward status. Each school and district in a reward status was provided reward dollars ranging between $2310 and $1155 per staff member (amount depended on the reward level of school/district). Reward checks were distributed on or about February 20, 1997.

For schools and districts whose performance judgment were designated as in decline or crisis (their final accountability growth index was less than their baseline) were provided the assistance associated with failing to meet improvement goals in the second accountability cycle. This start of this assistance was implemented on or about October 18, 1997 (the date scores were received).

District reports were issued at the same time as school reports, and differ from school reports only in terms of being based on all students in the district rather than all students in a school. Thus, district scores at a given accountability grade will not necessarily be equal to the weighted sum of the district's school's scores at that accountability grade. The inclusion of scores (and/or non-cognitive indicator results) from students who attend classes in a special learning environment, or who were not assigned to a reporting school, could alter district results.

School, district, and statewide-level results were released to schools, districts and media on October 18, 1997. These results were embargoed until October 23, 1997. The embargo allowed schools and districts to review their results and communicate these results to faculty and staff prior to their release to the general public. Concurrently, individual student results were also provided to schools, both in a summary format and on a report intended for distribution to the parents/guardians of each student who took a KIRIS test. Test result materials sent to schools and districts included:

- Individual Student Reports
- A Student Listing
- Item Level Report
- Assessment Curriculum Report
- Accountability Report

The remainder of this chapter describes in detail the final Cycle 2 KIRIS reports issued following the 1995-96 school year.

FOCUS GROUPS

To maximize the utility of any reports sent to either schools or parents various groups were asked to participate in numerous focus meetings. These meeting included parents, teachers, principals, public officials, and concerned citizens from across the state. Typically, each of these meetings included 10 to 25 people from one or more of the following groups: Louisville PTA, Cabinet for Human Resource (CHR) parent group, KIRIS Elementary School Principals advisory Committee, KIRIS Middle School Principals Advisory Committee, KIRIS High School Principals Advisory Committee, District Assessment Coordinators (DAC), and the Prichard Committee. At each of these meetings, the latest enhanced version of various reports were presented for discussion. Ideas gathered from each meeting would be presented to the next focus group for their recommendations. By completion of all the focus groups, numerous additions and refinements were incorporated into each report sent to schools and parents.

INDIVIDUAL STUDENT REPORTS

Two copies of each student's Individual Student Report (Appendix 10-1) were sent to schools. One report was forwarded to students' parents/guardians while the other report is for school use. These reports presented each student's principal performance level (Novice, Apprentice, Proficient, or Distinguished) in as many as six subject areas (reading, mathematics, science, social studies, writing portfolio, and mathematics portfolio) depending on the grade tested. Moreover, the individual student report also provided the student with information about their performance not found in any other report presented to schools. The principal performance levels of Novice and Apprentice were further divided into, Low, Middle, High categories. This additional parsing of both the Novice and Apprentice performance levels provides the student with a more precise idea of where his/her achievement was in relation to the next principal performance level.

The Individual Student Reports depicted the percentage of Kentucky students scoring in each of these performance levels for each of the subject areas at the student's grade. Each student's Kentucky percentile rank was given in four subject areas of reading, mathematics, science, and social studies, and for the composite of these subject areas. Numeric, as well as a visual representation of error bands was provided for the percentile rank associated with each subject area. For the composite percentile ranking, only a numeric indication of error[1] was provided.

Schools employed a variety of methods to transmit individual student data to the students' parents/guardians. Some schools simply sent the Individual Student Report to parents or guardians, some enclosed letters explaining results, while others asked parents/guardians to attend conferences, at which time results were explained in detail.

---

[1] Composite error bands across content areas for summed scores were based, in 1995-96, on the sum of the student-level subject area error estimates (after equation 23 of Feldt & Brennan, 1989, page 118). A different approach was used in calculating the error band around 1994-95 composite scores, but both expressed error in terms of the percentile rank corresponding to the score one standard error above and below the student's obtained percentile rank.

STUDENT LISTING

The Student Listing (Appendix 10-2) reports contained information about all students tested or accountable to a particular school by grade level. There are several student accountability types reported on the student listing. These student accountability situations include:

- Students tested and accountable at this school
- Students tested but accountable at another school
- Students tested at another school but accountable to this school
- Students tested with an Alternate Portfolio
- Students tested but exempt from accountability
- Students not tested but exempt from accountability

This listing reports each student name and their computer generated identification number. Along with student identification, each student's performance level and Kentucky achievement percentile in each of four open response content areas tested (reading, mathematics, science, and social studies) for the assessment grade was presented. If a writing or mathematics portfolio was required for assessment, the performance level as well as the instructional analysis for the writing portfolio and the core concepts analysis for the mathematics portfolio were presented.

Schools used the above information to not only review individual student achievement but also to make sure there was an accurate accounting of students accountable to the school. Scores obtained by students who were exempt from testing, according to Department of Education policy, are not aggregated into the school's total accountability score. However, these students are presented in student listing, so the school may identify any inconsistency with their records. Scores obtained by students in other accountability situations, (noted above) are also presented, so the school will be able to know where each accountable student was tested, and what scores were obtained to compute the school's index. Verification at the student level is an important check that each school performs.

ITEM LEVEL REPORT

Much like the Student Listing Report, the Item Level Report (Appendix 10-3) gives each student's name and computer generated identification number. However, unlike the Student Listing report the item level report provides detailed information about each student's response to each open-response question and the on-demand writing prompt. Each student's answers to open-response questions were evaluated on a five point, 0-4 scale. Below, is the non-grade or item specific scoring guide, which is used as a framework for grade and item specific scoring. This scoring framework is defined as follows:

Blank	The score of "blank" indicates a non-response. The student made no attempt to answer the question; the answer space was blank.
0	The score point of "0" indicates that a student's answer demonstrated one of two properties. It can mean the student's answer was totally incorrect, or it can mean the student's answer was off-topic, i.e., had nothing to do with the question, including irrelevant remarks.
1	The score point of "1" indicates that a student's answer demonstrated a minimal understanding of the

question. The student's response addressed the question but showed little knowledge about the topic. The student did not develop a complete answer, and answered only a small portion of the question.

2 The score point of "2" indicates that a student's answer demonstrated understanding of some of the important components of the question. This understanding was clearly communicated. However, the student's response also demonstrated some gaps in the student's conceptual understanding of the question.

3 The score point of "3" indicates that a student's answer demonstrated an understanding of most of the important components of the question. This understanding was clearly communicated. Moreover, the student's response also demonstrated an understanding of the major concepts even though some minor ideas or details were either overlooked or misunderstood.

4 The score point of "4" indicates that a student's answer demonstrated understanding of all of the important components of the question. This understanding was clearly communicated. The student demonstrated in-depth understanding of the relevant concepts and/or processes. Where appropriate, the student chose the more efficient and/or sophisticated process. Where appropriate, the student offered insightful interpretations or extensions (generalizations, applications, analogies).

The student responses to both the common (questions common to all students) and matrix sampled questions (questions that differ by test form for all students) are reported. The student's performance level for reading, math, science, social studies and on-demand writing are also indicated. To aid the school's assessment of student performance, the school, district, and state means are provided for each common test item.

ASSESSMENT CURRICULUM REPORTS

The Assessment Curriculum Report (Appendix 10-4) aggregates student level information into either the school or district level. The report contains the following information (excepting grade levels where a particular component, such as math portfolio, is not administered):

- Introduction
- Part I: Academic Results
    - Overall Summary of Results
    - Reading Results
    - Writing Portfolio Results
    - Mathematics Results
    - Mathematics Portfolio Results
    - Science Results
    - Social Studies Results
- Part II: Student Questionnaire Results
- Part III: Subgroup and Selected Analysis
- Part IV: District and State Summary Data

INTRODUCTION. The Assessment Curriculum Report introduction provides the reader with an overview of the contents of the report. It furnishes the background for the various parts of the report with regard to grade specific content areas. The introduction reviews the expectations that all "schools shall expect a high level of achievement of all students." It also describes the exemptions to that standard in the case of a.) Foreign exchange students, b.) Medical exemptions, and c.) Limited-English speakers.

OVERALL SUMMARY OF RESULTS. The Overall Summary of Results reports the most recent year's (1995-96) academic results by grade level assessed. The students scores have been aggregated by school (or by district for the district report). The school's/district's performance levels by content areas tested are reported. The table provides both the number and percentage of students in either the Novice, Apprentice, Proficient, or Distinguished, levels and Non-tested students by content area. (Non-tested students are students who did not respond to any question within a specific content area; these non-responsive students are considered Novices in that specific content area for assessment purposes.) The reporting of Non-tested (non-responsive) students provides schools with supplementary information in addition to student performance level. Also, the number and percentage of alternate portfolio students and non-participating students by exemption (foreign exchange, medical exemptions, and limited-English speakers) are reported.

READING RESULTS. The on-demand performance results for the content area of reading are reported in this section of Part I: Academic Results. For each year of Cycle 2 the number and percentage of Novice, Apprentice, Proficient and Distinguished responses are tabulated. Within the reading content area, defined by an Academic Expectation, are skills and strategies that students will need to use as they work on the KIRIS assessments. Students' ability to make sense of a wide variety of materials including informational texts, literary texts, persuasive texts, and practical reading materials are important skills. These four subcategories of reading are also reported within this section. For the current year, the school's/district's Novice, Apprentice, Proficient and Distinguished percentages are given for each of these reading subcategories.

WRITING PORTFOLIO RESULTS. The portfolio performance results for the content area of writing are reported in this section of Part I: Academic Results. For each year of Cycle 2 the number and percentage of Novice, Apprentice, Proficient and Distinguished portfolios are tabulated. Many of goals of the writing content area, defined by an Academic Expectation, can be demonstrated using writing portfolio assessment. The writing portfolio scoring guide, used in the holistic scoring of the writing portfolio, also provides teachers with an instructional analysis useful in annotating student portfolios to indicate instructional strengths. The following areas are used to indicate instructional strengths: 1.) Establishing focused authentic **Purpose**, 2.) Writing for authentic **Audiences**, situations, 3.) Employing a suitable **Voice and/or Tone**, 4.) **Developing Ideas** relevant to the purpose, 5.) **Supporting** ideas with elaborated, relevant **Details**, 6.) **Organizing** ideas logically, 7.) Using effective **Transitions**, 8.) Constructing effective and/or correct **Sentences**, 9.) Using **Language** effectively and/or correctly, 10.) **Editing** for correctness.

For the current year, the number and percentage of portfolios showing instructional strengths are tabulated for the school/district. Schools can examine their instructional strengths using this table. This review can assist schools in improving writing and learning.

MATHEMATICS RESULTS. The on-demand performance results for the content area of mathematics are reported in this section of Part I: Academic Results. For each year of Cycle 2 the number and percentage of Novice, Apprentice, Proficient and Distinguished responses are tabulated. Within the mathematics content area, defined by academic expectations, are a common core of important mathematics skills that students will need to use as they work on the KIRIS assessments. There are seven identified core concepts which are:

- Comprehensive understanding of **Number** and counting processes.
- Recognizing **Mathematical Procedures** as part of the study of patterns, to find solutions to numerical questions. The use of creative design and logical thinking to help solve problems.
- Classifying and using **Space and Dimensionality** concepts in the solving of spatial relationship problems.
- Mastering the concepts of **Measurement** in the measuring of appropriate attributed (in liquid, solid, time, place) units, and successfully applying these concepts to problems in daily life and business.
- Using mathematics to help measure and quantify **Change** in patterns and situations that apply to business and daily life.
- Analyzing and understanding the **Mathematical Structure** of one mathematical core concept in light of other core mathematical concepts. The application of the understanding of mathematical structure relationships to inventive problem solving.
- Mastering the ability to access, interpret, use and disperse **Data** effectively.

These seven core concepts of mathematics are collapsed for on-demand reporting within this section. The three collapsed reporting categories are: 1.) Number, Mathematical Procedures, and Mathematical Structure, 2.) Space and Dimensionality and Measurement, and 3.) Change and Data. For the current year, the school's/district's Novice, Apprentice, Proficient and Distinguished percentages are given for each of these three on-demand mathematics reporting categories.

MATHEMATICS PORTFOLIO RESULTS. The portfolio performance results for the content area of mathematics are reported in this section of Part I: Academic Results. For each year of Cycle 2 the number and percentage of Novice, Apprentice, Proficient and Distinguished responses are tabulated. Many of goals of the mathematics content area can be demonstrated using mathematics portfolio assessment. The same seven identified core concepts of Number, Mathematical Procedures, Space and Dimensionality, Measurement, Change, Mathematical Structure, and Data, used in the on-demand portion of the KIRIS, are also used in mathematics portfolio assessment. The mathematics portfolio scoring manual provides teachers with procedures for assigning a holistic performance rating based on an overall view of the full contents of a portfolio.

The mathematics portfolio results section also provides the current year summary of the mathematics portfolio core concept data analysis. This analysis provides teachers with a broad view of how well students understood the core concepts presented in the portfolios. The seven mathematics core concepts are analyzed with regard to student's basic understanding, understanding, broad understanding, or comprehensive understanding of these concepts. This analysis provides both the number and percentage of students that are placed into each of these four categories.

SCIENCE RESULTS. The on-demand performance results for the content area of science are reported in this section of Part I: Academic Results. For each year of Cycle 2 the number and percentage of Novice, Apprentice, Proficient and Distinguished responses are tabulated. Within the science content area, defined by academic expectations, are a common set of themes that permeate the sciences, be it life sciences, earth and space sciences, or the physical sciences. Identified are six themes that are important to science assessment. These six science themes are:

- Process and **Nature of Scientific Activity**
- **Patterns** (e.g. trends, cycles in the sciences)
- **Systems** and subsystems - structures and functions
- Scientific **Models and Scales**
- **Constancy** and equilibrium in nature
- **Evolutionary** change over time

These six concepts of science are collapsed for on-demand reporting within this section. The three collapsed reporting categories are: 1.) Nature of Scientific Activity, and Models and Scales, 2.) Patterns, Constancy, and Evolution, and 3.) Systems. For the current year, the school's/district's Novice, Apprentice, Proficient and Distinguished percentages are given for each of these three on-demand science reporting categories.

SOCIAL STUDIES RESULTS. The on-demand performance results for the content area of social studies are reported in this section of Part I: Academic Results. For each year of Cycle 2 the number and percentage of Novice, Apprentice, Proficient and Distinguished responses are tabulated. Social studies content area, defined by academic expectations, are concentrated on five areas. These five social studies assessment areas are:

- History
- Government (including civics, and citizenship)
- Culture
- Geography
- Economics

These five content areas of social studies are further integrated into three reporting categories for on-demand reporting within this section. These reporting categories are: 1.) Democratic Principles and Political Systems, 2.) Social Systems, Economic Systems, and Cultural Diversity, and 3.) Relation of Geography to Human Activity and Historical Perspective. For the current year, the school's/district's Novice, Apprentice, Proficient and Distinguished percentages are given for each of these three on-demand social studies reporting categories.

STUDENT QUESTIONNAIRE RESULTS. All students in the accountability system were administered a questionnaire at the end of open-response testing. Questions administered varied slightly from grade to grade. Both the number and percentage of student responses to each question were reported. The questions in the fourth and eighth grades included length of Kentucky residency and current school attendance, after school activities to include the amount of homework, nonacademic reading, and television, the amount of school course coverage in relation to test coverage, the amount of participation in various activities including: groups, projects, information retrieval, oral reports, use of calculator, free choice reading, mathematics problem solving, and use of mathematics manipulatives. For twelfth grade students, some fourth and eighth student questions were augmented and replaced. Additional questions include the amount of useful teacher feedback given on homework, the amount and concordance of part-time work with career goals, absenteeism in a given month, vocational plans, the number of courses by subject area that will be completed at high school graduation, current academic performance in school, and current English and mathematics academic level course enrollment.

SUBGROUP AND SELECTED ANALYSES. The subgroup analysis reports the number and percentage of Novice, Apprentice, Proficient and Distinguished responses across all content areas by three subgroups. These content areas include reading, mathematics, science, social studies, arts and humanities, practical living and vocational studies, writing portfolio, and mathematics portfolio. The three subgroups are:

- Gender, (male/female)
- Race, (white/non-white)
- Title 1, (targeted assistance/non-targeted assistance schools)

The subgroup identification process only considered data scanned from student answer documents. To protect anonymity of respondents, no data are reported if a category includes fewer the 10 students. The analyses do not include students who are participating in alternate portfolios.

DISTRICT AND STATE SUMMARY DATA. district and state summaries provide schools with referents to each of the content areas tested with KIRIS. The percentage of students in the district and state at each performance level within each content area are given to help schools analyze their performance.

ACCOUNTABILITY REPORT

The purpose of the KIRIS Accountability Cycle 2 report (Appendix 10-5) is to provide a school or district a single two page report that has all their KIRIS accountability information. The report provides the accountability index scores by which schools were evaluated for the second accountability cycle. Using these scores a school's/district's performance judgment was made.

The first page of the accountability report captures some of the same elements found in the Assessment Curriculum Report. For each year of the four year second accountability cycle (1992-93, 1993-94, 1994-95, and 1995-96) the percentage of Novice, Apprentice, Proficient and Distinguished at the school or district level are reported by content area. The content areas reported are: reading, mathematics, science, social studies, arts & humanities, practical living and vocational studies, and writing portfolio. If a content area is represented by various types of testing, (i.e. open-response, and portfolio) these values and their weighted totals are also given.

The second page of the accountability report provides schools/districts with their academic, noncognitive , and accountability indices. There is an academic index for each content area. A content area specific academic index is computed by multiplying the percentage Novice, Apprentice, Proficient and Distinguished in that content area by 0, 40, 100 and 140, respectively and summing the products. The weighted academic indices are combined with a weighted noncognitive index to form the accountability index. The mathematics, reading, science, social studies, and writing indices are weighted 14% each, Arts and Humanities, and Practical Living and Vocational Studies are weighted 7% each, and the noncognitive index is weighted 16%. For schools also interested in reporting only the academic portion of the accountability index, the total academic index is also provided. This index uses only content area weights used with the accountability index to compute the academic index; no noncognitive information is included. However, since the combined cognitive weights only equal 84% the sum of the products must be divided by .84 to place the resulting total academic index on the same metric as the accountability index.

To provide a school/district with a performance judgment (Reward, Successful, Successful Year 2, Improving, Improving Category 2, Decline, or Crisis) the school's/district's Baseline Index, Improvement Goal, and Combined Growth Index must be computed. The accountability report provides all of these indices. The Baseline index, which is the weighted average of the accountability indices for the first two years of the accountability cycle, is used to calculate the amount of growth required of the school/district during the cycle. This growth is added to the baseline to provide the Improvement Goal. The average Accountability Growth Index which is the weighted average of the accountability indices for the last two years of the accountability cycle, is then compared to the Improvement Goal to produce a performance judgment. Each accountability report has a tailored message indicating the school's/district's performance judgment.

These reports are further described in the *Interpretive Guide*, published in October 1996, which accompanied the reports.

CHAPTER 11
STATEWIDE SUMMARY OF RESULTS

Table 11-1 summarizes the final accountability index results for the second Accountability Cycle[1]. Tables 11-2 presents the same analysis for all schools by region, while Table 11-3 provides this information by content area, and table 11-4 presents results for non-cognitive indicators. The data in Tables 11-3 and 11-4 are based on weighted averages of school results.

TABLE 11-1
STATEWIDE ACCOUNTABILITY INDEX SCORES BY GRADE

Grade Level	1992-93 Index	1993-94 Index	1992-94 Baseline Index	1994-96 Improvement Goal	1994-95 Index	1995-96 Index	1994-96 Accountability Growth Index
Grade 4	35.7	40.9	38.3	44.5	47.1	45.2	46.2
Grade 8	37.4	41.8	39.6	45.6	44.5	41.0	42.8
Grade 11/12	34.9	43.3	39.0	45.1	44.6	43.3	44.0

ACCOUNTABILITY INDEX SCORES

Statewide performance showed some impressive gains on the 1994-96 Accountability Growth Index scale during the second accountability cycle. Accountability indices for the state as a whole averaged across the grades shown in Table 11-1, increased from the Baseline Index (the weighted average of the 1992-93 and the 1993-94 index) to the Accountability Growth Index (the weighted average of the 1994-95 and the 1995-96 index). The statewide Accountability Growth Index of 44.0 showed a 5.3 point gain from the statewide Baseline Index of 39.0. However, the statewide Accountability Growth Index did not reach the statewide Improvement Goal of 45.1 points, (6.1 points above the statewide baseline). Never the less, the 5.3 point obtained gain did constitute an 87 percent achievement of the state's improvement goal[2].

The average score for grade 4 rose from its baseline of 38.3 to its Accountability Growth Index of 46.2, a 7.9 point rise. At grade 8, the average index rose from a baseline of 39.6 to the 1994-96 average Accountability Growth Index score of 42.8, for a total gain of 3.2 points. At grade 11/12, the overall rise was similar to grade 4, with a 1992-94 Baseline Index of 39.0 and a 1994-96 average Accountability Growth Index of 44.0, a 5.0 point gain. Moreover, grade 4 showed the largest percentage of goal achieved of any of the three grade levels. Grade 4's 7.9 point rise showed a 127 percent achievement of their desired gain of 6.2 points. Grade 8 showed the smallest percent of goal achieved of any grade level. Grade 8's 3.2 point rise showed only a 53 percent achievement of their desired gain of 6.0 points. Grade 11/12 was a middle performer showing a 5 point gain from the desired 6.1 points. This gain translated to an 82 percent achievement of their desired improvement for grade 11/12.

---

[1] Successful appeals may modify some of these data.

[2] A convenient way to evaluate state achievement is to calculate the percentage of the state goal obtained. Ideally, the state should achieve 100% or more of its goal. However, during the second accountability cycle the state achieved 87 percent of its overall goal, i.e., increased its score form the baseline score by 5.3 points of the necessary 6.1 points to meet its Improvement Goal, thus achieving 87 percent of its 6.1 point goal.

TABLE 11-2
STATEWIDE ACCOUNTABILITY INDEX SCORES, BY GRADE AND REGION

Grade Level	Region Number	1992-93 Index	1993-94 Index	1992-94 Baseline Index	1994-96 Improvement Goal	1994-95 Index	1995-96 Index	1994-96 Accountability Growth Index
Grade 4	Region One	35.6	40.7	38.2	44.4	47.7	43.8	45.8
	Region Two	35.7	41.1	38.4	44.6	47.5	47.0	47.3
	Region Three	36.0	40.4	38.2	44.4	47.6	42.7	45.1
	Region Four	36.9	42.9	39.9	45.9	48.3	46.4	47.4
	Region Five	36.8	42.5	39.6	45.6	48.5	47.6	48.1
	Region Six	34.5	39.7	37.1	43.4	45.0	44.1	44.6
	Region Seven	36.8	41.0	38.9	45.0	46.4	45.9	46.2
	Region Eight	32.8	37.7	35.2	41.7	44.0	42.7	43.4
Grade 8	Region One	38.7	41.9	40.3	46.3	45.5	41.6	43.6
	Region Two	38.6	42.7	40.7	46.6	45.8	41.8	43.8
	Region Three	35.2	39.8	37.5	43.8	40.6	36.2	38.5
	Region Four	39.2	43.3	41.3	47.2	46.2	42.6	44.4
	Region Five	39.9	45.2	42.6	48.3	47.4	43.2	45.3
	Region Six	35.1	39.3	37.2	43.5	43.3	40.3	41.8
	Region Seven	37.4	41.1	39.3	45.4	43.6	41.6	42.6
	Region Eight	34.7	40.3	37.4	43.7	42.4	40.5	41.4
Grade 11/12	Region One	34.0	44.2	39.0	45.1	47.2	45.2	46.2
	Region Two	36.5	42.8	39.6	45.6	45.0	43.9	44.4
	Region Three	33.5	42.8	38.1	44.3	42.5	40.9	41.7
	Region Four	37.2	48.2	42.5	48.3	46.5	46.2	46.3
	Region Five	37.0	45.3	41.0	46.9	46.3	45.4	45.8
	Region Six	33.6	41.4	37.4	43.7	43.6	40.8	42.2
	Region Seven	34.7	41.2	37.9	44.1	44.3	42.2	43.3
	Region Eight	30.6	38.6	34.6	41.1	39.8	40.3	40.1

Table 11-2 subdivides the statewide accountability score by grade and region. All regional grade 4 accountability indices recorded scores over 100 percent of their improvement goal. However, region 3 showed the lowest percentage of goal achieved showing 111 percent. Grade 8, across all regions, showed an entirely different pattern. No region achieved 100 percent of its improvement goal. The majority of the regions were in the 50 to 60 percent range, with region 6 obtaining 73 percent and region 3 showing only 16 percent of improvement goal obtained. Grade 11/12 showed some improvement over grade 8 but was not close to showing the improvement gains noted by grade 4. Only region 1 reported a percentage improvement goal above 100 percent, with a percent improvement of 118 percent. Most of grade 8 improvement percentages were situated between 60 to 80 percent, with region 3 being the lowest at 58 percent.

TABLE 11-3
CONTENT AREA INDEX SCORES BY GRADE AND YEAR

Grade Level	Year	Reading	Math	Science	Social Studies	Arts & Humanities	Practical Living	Writing	Accountability* Index
Grade 4	1992-93	32.4	22.3	18.2	27.6	8.4	19.2	31.0	35.7
	1993-94	39.8	34.2	22.4	32.7	9.2	21.7	37.6	40.9
	1994-95	55.2	41.7	32.4	38.8	11.8	24.0	40.3	47.1
	1995-96	56.5	38.8	29.8	35.9	6.2	18.7	39.3	45.2
Grade 8	1992-93	38.4	22.8	19.9	28.9	20.3	16.0	28.0	37.4
	1993-94	46.8	31.4	20.7	38.1	19.6	19.0	31.4	41.8
	1994-95	45.4	40.0	24.1	42.9	23.1	22.2	31.8	44.5
	1995-96	45.5	40.6	18.6	35.9	18.9	8.8	27.5	41.0
Grade 11/12	1992-93	20.2	22.2	27.1	23.4	14.3	11.8	41.1	34.9
	1993-94	34.2	30.4	36.8	41.2	23.8	24.9	39.6	43.3
	1994-95	33.8	36.2	40.9	40.1	17.2	25.3	38.7	44.6
	1995-96	33.3	36.4	40.7	34.5	16.8	19.7	38.9	43.3

*Includes Non-cognitive Index

GRADE 4. Index scores ian all seven content areas - - reading, math, science, social studies, arts & humanities, practical living & vocational studies, and writing - - rose from 1992-93 to 1994-95, most climbed substantially. Reading rose 22.8 points, math rose 18.4, science rose 14.2 points, social studies rose 11.2 points, arts and humanities rose 3.4 points, practical living & vocational studies rose 4.8 points while writing rose 9.3 points. However, this pattern did not continue for the 1995-96 accountability year. Only reading rose moderately at 1.3 points, all other content areas decreased between 1.0 and 5.3 points. The largest drops in scores occurred in two content areas of arts & humanities, and practical living and vocational studies. A minor change was observed in non-cognitive indicators. A drop of 0.2 points occurred from 1992-93 to 1995-96.

GRADE 8. Much like the grade 4 results for the second accountability cycle, grade 8 scores rose from 1992-93 to 1994-95, most climbed substantially. Reading rose 7.0 points, math rose 17.2, science rose 4.2 points, social studies rose 14.0 points, arts and humanities rose 2.8 points, practical living & vocational studies rose 6.2 points while writing rose 3.8 points. Much like 4th grade, this pattern did not continue for the 1995-96 school year. Only writing and math rose slightly, with a 0.1 and 0.6 point gain respectively. All other content areas decreased between 4.3 and 7.0 points. A minor change was observed in non-cognitive indicators. A drop of 0.3 points occurred from 1992-93 to 1995-96.

TABLE 11-4
STATEWIDE NON-COGNITIVE INDICATORS BY GRADE AND YEAR

Grade Level	Year	Attendance Rate	Retention Rate	Dropout Rate	Transition to Adult Life	Overall Non-cognitive Index
Grade 4	1992-93	95.49	0.68	n/a	n/a	96.3
	1993-94	95.64	0.64	n/a	n/a	96.4
	1994-95*	95.64	0.64	n/a	n/a	96.4
	1995-96	95.35	1.06	n/a	n/a	96.1
Grade 8	1992-93	95.05	1.44	0.75	n/a	97.3
	1993-94	95.10	1.64	0.82	n/a	97.2
	1994-95*	95.10	1.64	0.82	n/a	97.2
	1995-96	94.75	1.74	0.80	n/a	97.0
Grade 11/12	1992-93	93.32	5.12	5.06	92.59	93.7
	1993-94	93.13	5.33	5.14	93.99	94.2
	1994-95*	93.13	5.33	5.14	93.99	94.2
	1995-96	92.62	6.37	5.43	93.64	93.8

* Beginning 1994-95, the non-cognitive data was lagged by one year.

GRADE 11/12. Grade 11/12 showed a different pattern from grades 4 and 8. All but one grade 11/12 score rose from 1992-93 to 1993-94, most climbed substantially. Reading rose 14.0 points, math rose 8.2, science rose 9.7 points, social studies rose 17.8 points, arts and humanities rose 9.5 points, practical living & vocational studies rose 13.1 points while writing decreased by 1.5 points. However, not all of the 1994-95 scores rose. While math rose 5.8 points and science rose 4.1 points, there was a general flatting of reading, social studies, practical living and vocational studies, and writing scores; arts and humanities was alone in its -6.6 point drop. In 1995-96, five of seven content area scores were basically flat relative to 1994-95 scores. Reading changed -0.4, math changed 0.2, science changed -0.2, arts and humanities changed -0.4, and writing changed 0.2 points. However, social studies, and practical living and vocational studies showed much greater decreases. Both social studies, and practical living and vocational studies indices decreased by 5.6 points. A minor increase of 0.1 points in the non-cognitive index was observed from 1992-93 to 1995-96. The increased rate of successful transition to adult life, as judged by graduating students working, joining the military, or pursuing post-secondary education in 1993-94 (which was lagged for 1994-95) was unable to compensate for the decline of the attendance rate and an increase in the retention and dropout rates for grade 11/12 found in 1995-96.

SECTION IV
GENERALIZABILITY

Section 4 examines issues related to test score validity in a broad sense. Chapter 13 presents reliability estimates for several components of the test. Chapter 14 describes existing validity-related evidence for KIRIS. Chapter 15 explores consequential validity for the assessment, including differences in performance at the student level by gender and ethnicity.

CHAPTER 12
PORTFOLIO ASSESSMENT: SCORING AND STUDENT PERFORMANCE

THE PLACE OF PORTFOLIO ASSESSMENT IN KIRIS

Portfolios occupy a key place in KIRIS, both as a means of assessment that directly taps student work in classrooms, and as a means for supporting educational improvement in classrooms, schools, and districts. Since the contents of the portfolios arise from students' classroom work, the portfolio is the assessment component that most clearly reflects local curriculum and instruction. In concept, portfolios are developed by students over long periods of time--months and perhaps years. Because students have had the opportunity to revise their portfolio entries--with support and feedback from teachers and peers-- the assessment portfolio may reasonably be viewed as the student's "best work."

In many respects portfolios are the portion of the KIRIS assessment that most directly and comprehensively supports educational reform because of the strong connection to students' classroom experiences and strong involvement of teachers. For that reason, the KIRIS portfolio activities include extensive professional development opportunities which local schools can employ as a powerful means to support teachers' professional development and school improvement. Both writing and mathematics employ a trainer-of-trainers model to deliver scoring training throughout the state. In addition, regional consultants are available to provide professional development and informal teacher support throughout the year.

Writing portfolios were fully implemented in the first KIRIS accountability cycle, and continued in Cycle 2; mathematics portfolios and alternate portfolios were introduced during the second accountability cycle. All KIRIS portfolios are scored locally to allow each school to observe all of the information included in the portfolios, information that goes well beyond the scoring criteria. Reliance on local scoring requires training and practice as well as alignment between portfolio requirements and local instruction. The portfolio development and scoring process also assumes considerable content knowledge of teachers. Local scoring provides a wholly different model of teacher responsibility and involvement in the accountability system than centrally scored or multiple-choice tests. Although external scoring can provide summary data in the form of scores and standardized notations (see description of scoring analysis), only local portfolio scoring allows discussion of the best ways to modify instruction based on assessment data that directly reflect classroom practices. Extensive professional development is provided throughout the state to support scoring accuracy and the alignment of instruction with the portfolio assessment criteria.

All KIRIS portfolio assessments are standardized in the following ways. 1) All require the use of a scoring guide accompanied by several benchmarks or exemplars for each score point. 2) Portfolio content requirements prescribe the number and types of required entries. Portfolios that are incomplete are scored as Novice (zero points). 3) The state provides every school district with complete scoring training materials for each accountability grade, including detailed rules for portfolio preparation to assure that the work in each portfolio has been completed by the student.

MONITORING THE SYSTEM

Portfolio development and scoring are monitored in several ways. Early in the year, a Code of Ethics is distributed to all schools describing the limits on a teacher's comments or modifications of a student's portfolio entries. When the school assigned portfolio scores are submitted to KDE, the principal is required to submit a signed assurance statement confirming that appropriate portfolio development practices were observed. Accusations or complaints of inappropriate practices are investigated by KDE and penalties applied, if warranted. In addition, each student portfolio includes a statement signed by the student attesting that all portfolio entries were completed by the student. If plagiarism is discovered at local scoring or during an audit, the entry is "removed," making the portfolio incomplete. Incomplete portfolios receive a score of zero.

The public accountability and high stakes associated with the assessment system necessitate continued monitoring of portfolio development and scoring. Two different types of activities are used to provide schools with feedback on their scoring effectiveness, the scoring audit and the scoring analysis. Formal studies of local scoring accuracy are referred to as an "audit." The audit has several purposes: 1) to monitor accuracy of scoring throughout the system in order to plan statewide training and allocation of resources; 2) to correct inaccurate scores assigned locally; and 3) to verify exceptional score gains. KDE defines the sample of schools, usually in a manner that allows allow the results to be generalized beyond the group of participating schools. All schools that are selected must participate; they are required to submit all portfolios for rescoring; and locally assigned portfolio scores and the resulting Cognitive Index are changed as a result of the audit.

The other type of activity is referred to as "scoring and analysis." Scoring Analysis sessions are designed to provide schools with information about their scoring accuracy without the risk of changes to their school accountability scores. Many of the same scoring procedures that are used in an audit are employed, but neither scores nor the accountability index are changed. The emphasis is on providing commentary, based on examination of a school set of portfolios, that will help schools plan future instruction and professional development. School participation is voluntary.

After five years of portfolio assessment, two main issues continue to surround the use of portfolios in Kentucky's assessment system: the level of scoring accuracy achieved by Kentucky teachers and the impact of portfolios on instructional practice. The following sections present the rationale for the design of each portfolio, information about the scoring reliability and instructional impact of each during Cycle 2 (the 1992-93 school year through 1995-96),[1] and a discussion of related issues.

---

[1] Information about activities occurring prior to 1992-93 may be found in the technical manual for the corresponding year.

THE KIRIS WRITING PORTFOLIO

RATIONALE AND DESIGN. The Kentucky Writing Portfolio assesses student writing directly (at Grades 4, 8, and 12) by examining a collection of a student's written products. The structure of the writing portfolio and the holistic scoring guide encourage teachers to provide instruction focused on teaching students to communicate effectively and to provide grammar, punctuation, and spelling instruction through these authentic writing experiences.

The portfolio was originally designed by a committee of Kentucky English/Language Arts educators. This committee discussed the traditional writing experience of Kentucky students and found that most instruction had focused on isolated grammar and very confined writing experiences (i.e., reports, essays, research papers). Using the writing Academic Expectation as their guide (that all students should write for multiple purposes in multiple forms for a variety of audiences), the committee structured the contents of the portfolio to include broad categories of writing that consciously excluded reports, academic essays, and research papers. Instead, the committee created a structure that required other types of writing.

- Personal experience writing
- Imaginative writing
- Reflective writing
- Transactive writing for real-world purposes and audiences

In addition to this purposeful design of the portfolio contents, the criteria for assessment were selected and scoring tools designed with these instructional focus changes clearly in mind. While the committee believed that mastery and assessment of mechanics remained critical, they also identified several more critical criteria that had traditionally been less evident in writing instruction and assessment in Kentucky (e.g., organizational skills, idea development, and focus on purposes and audiences). Finally, the committee selected six main criteria for assessing the quality of student writing. These criteria are applied holistically producing a single final judgment for a complete portfolio.

- Purpose/Audience Awareness
- Idea Development/Support
- Organization
- Sentence Structure and Variety
- Language (Word Choice and Usage)
- Correctness (Spelling, Punctuation, and Capitalization)

The committee believed that these portfolio content requirements and assessment criteria would provide teachers with guidelines for more balanced writing instruction, consistent with the national movement toward more process-centered instruction.

PROFESSIONAL DEVELOPMENT

The Kentucky Writing Program (KWP) supports a wide variety of professional development experiences including portfolio scoring training as well as workshops and consultation focused primarily on classroom strategies for developing student writing skills. Since the introduction of the writing portfolio assessment, the KWP has supported classroom teachers through a tiered training system that relies on a design committee to train local trainers; the local trainers then deliver portfolio development and scoring strategies to the other teachers in their school. Each year, these local trainers receive two rounds of professional development: one focused on the generation of portfolio entries and one focused on scoring to state standards. These sessions are augmented by print materials, video training provided through state-wide educational television, and local level workshops provided by the Kentucky Writing Program regional consultants.

In addition, Regional Writing Consultants work with local districts and schools upon request to provide individually tailored professional development experiences focused on a variety of writing portfolio related topics such as portfolio analysis, technical writing, personal experience writing, reflective writing, writing across the curriculum, development of writing workshop classrooms, and designing appropriate assignments focused on real-world purposes and audiences.

1996 WRITING PORTFOLIO SCORING AUDIT[2]

The 1996 Writing Portfolio scoring audit was the second to be carried out. The results and the consequences for schools were very different from the previous audit. The first writing portfolio scoring audit was held in 1993 (for details see Cycle I Technical Report). By legislative directive, the 1993 Writing Audit allowed schools the choice of keeping their original scores or accepting revised scores based on the generally lower audit results. Most schools chose to use the scores that they had assigned to the portfolios to compute both their Writing Cognitive Index and the KIRIS Accountability index.

In 1996, a second audit was conducted to monitor state-wide scoring patterns and adjust scores for those schools found to be scoring portfolios inaccurately. The audit results were reported to individual schools; the audit scores were used to adjust the writing cognitive index of all audited schools; and overall, local scoring was observed to be much more accurate than in 1993.

The 1996 audit was carried out by the scoring contractor with a small proportion of the audit portfolios rescored by experienced Kentucky teachers to assure that the contractor had applied the scoring standards correctly. Audit scorers were trained in use of the Kentucky Scoring Guides, required to qualify by scoring a set of portfolios to a predetermined level of accuracy, and monitored continuously for scoring accuracy. Any audit scorer who failed to maintain the required level of scoring accuracy was discharged.

---

[2] For complete procedures and full results of the audit, see *1996 Writing Portfolio Audit: Final Report*

One hundred schools were selected for auditing. (Because two of the selected schools had been closed, the final group of schools audited numbered 98.) Half, designated as the Random Schools, were selected at random, providing a sample of schools from which to infer statewide scoring accuracy rates. The other half, referred to as the Purposeful Schools, were chosen using a formula that identified schools with writing portfolio scores that were very high or low relative to test scores in other content areas.

The results of the audit are summarized in Table 12-1. Two measures of scoring accuracy are presented, the exact agreement between the portfolio scores assigned by the school and those assigned by the audit, and the magnitude of the difference in the Writing Cognitive Index (WCI), a scale of 0 to 140 points. The Purposeful group exhibited slightly lower exact agreement with the audit scores; seventy-three percent of their portfolio scores were confirmed by one or more audit readers. For the Random group, seventy-seven percent of the locally assigned scores were confirmed. Across schools, the mean change in the Writing Cognitive Index was nearly twice as large for the Purposeful group as for the Random group. The audit adjusted the WCI for the Purposeful group downward by nearly 11 points, while the change for the Random schools was a reduction of approximately 5 points. The Writing Cognitive Index is determined by assigning a score of 0 for all portfolios scored "Blank," "Incomplete," or "Novice." Values of 40, 100, and 140 points are assigned to portfolios rated as Apprentice, Proficient or Distinguished, respectively. The Writing Cognitive Index (WCI) is then computed as the arithmetic mean of all portfolio scores.

TABLE 12-1
SUMMARY OF 1996 AUDIT RESULTS

Sample	Original-Audit Score Agreement	Mean Change in Writing Cognitive Index
Purposeful	73%	-10.87
Random	77%	-5.09

It is important to note that scoring accuracy was not uniform across schools. Nearly 70% of the schools in the 1996 audit assigned scores in exact agreement with the audit for seventy percent or more of their portfolios. This is a remarkable change from 1993 when no schools scored portfolios at this level of accuracy. Table 12-2 provides comparison data based on the exact agreement between locally assigned scores and audit scores.

The first audit of writing portfolios, conducted in 1993, demonstrated that schools were significantly over-estimating student performance. On average, the audit-score WCI was nearly 40 points lower than the school-score WCI. The 1996 audit presents very different results. [3]

---

[3] It is important to note that the selection process used for this first auditing activity was solely purposeful (e.g., all schools were selected for auditing due to apparent discrepancies in score gains in writing relative to other content areas). Therefore, comparisons between this early audit and the recent audit (discussed below) can only be made within the pool of purposefully selected schools. (For more information about the 1993 audit see the Cycle I Technical Report.)

TABLE 12-2
1993 AND 1996 AUDIT: LEVELS OF EXACT AGREEMENT

Percentage of Exact Agreement	1993 Audit	1996 Audit		
	Number of schools (Purposeful)	Number of Schools (Purposeful)	Number of Schools (Random)	Number of Schools (Total)
90% or more	0	6	4	10
80 - 89%	0	12	21	33
70 - 79%	0	14	11	25
60 - 69%	1	10	8	18
50 - 59%	4	4	4	8
less than 49%	89	3	1	4

TABLE 12-3
COMPARISON OF 1993 AND 1996 WRITING PORTFOLIO
CHANGES TO WRITING COGNITIVE INDEX

	1992-93 Audit Results		1995-96 Audit Results			
	Purposeful Schools		Purposeful Schools		Random Schools	
	Original WCI*	Audited WCI (Difference)	Original WCI	Audited WCI (Difference)	Original WCI	Audited WCI (Difference)
Grade 4	64.0	19.6 (-44.4)	54.6	46.0 (-8.6)	40.4	35.2 (-5.2)
Grade 8	55.1	14.3 (-40.8)	26.3	12.9 (-13.4)	25.8	21.5 (-4.3)
Grade 12	64.1	31.1 (-33.0)	55.4	42.0 (-13.4)	29.4	23.9 (-5.5)
Total Schools	61.1	21.6 (-39.5)	44.0	33.1 (-10.9)	31.9	26.8 (-5.1)

*Writing Cognitive Index

The results for 1996 demonstrate substantial improvement in local scoring accuracy since 1993.

Together these two pieces of information, writing cognitive index changes and percents of exact agreement, indicate overall improvement from 1993 to 1996 in scoring accuracy for purposefully selected schools. While some schools still display significant scoring discrepancies, even those schools are demonstrating substantially higher agreement than similar schools in the 1993 audit.

STUDENT PERFORMANCE

The data presented above suggest that conclusions about changes in student performance between 1993 and 1996 must be tempered by recognition of the increased scoring accuracy during the same period. Table 12-4 shows the state mean score on the writing portfolio over the last four years. While there is little change in writing portfolio scores since 1993 in middle and high schools and only modest gains in elementary, the increasing accuracy in local scoring (reduction of overestimated scores) probably masks real gains in student performance.

TABLE 12-4
STATEWIDE WRITING PORTFOLIO MEAN SCORES

GRADE	1992-93	1993-94	1994-95	1995-96
4	31.0	37.6	40.3	39.3
8	28.0	31.4	31.7	27.5
12	41.1	39.6	38.6	38.9

USING WRITING PORTFOLIO SCORING ANALYSIS TO PROVIDE INSTRUCTIONAL FEEDBACK

During the summers of 1994 and 1995, all Kentucky schools had the opportunity to submit portfolios for Scoring Analysis. The purpose of portfolio Scoring Analysis is twofold: to provide a school with informal feedback on their scoring accuracy and to provide recommendations about instructional practices and future professional development based on a review of portfolio contents. Also, the Scoring Analysis session models an activity that schools can carry out independently. School participation was completely voluntary. Although we cannot generalize the results from this activity to draw firm conclusions about typical instructional practices statewide, the Scoring Analysis does provide information about widespread teacher understanding of best practices in writing instruction.

Scoring Analysis scorers were Kentucky teachers trained and qualified to score and analyze school sets of portfolios. Scorer accuracy was monitored throughout the session. These Kentucky educators studied school sets of 20-30 portfolios to score student work, and to evaluate instructional strengths on the basis of the writing products in those portfolios. Each school set of portfolios was evaluated by a panel of six readers which conferred to draft recommendations concerning future instructional focuses and professional development for the school.

RECOMMENDATIONS. The 1995 Scoring Analysis examined more than 10,000 portfolios. A team of six readers worked together to review and discuss instructional evidence present in the portfolios from a particular school. Feedback was provided by noting extraordinary performance on ten descriptors that are based on the six assessment criteria of the scoring guide. After discussing the school set of portfolios, the group reached consensus on which of ten possible descriptors should be checked as "Strength" or "Need." (A neutral result was not noted.) A single report was prepared for each school. Table 12-5 summarizes the results across all participating schools.

TABLE 12-5
INSTRUCTIONAL ANALYSIS STRENGTHS AND NEEDS
(1994-95 SCORING ANALYSIS)

Instructional Analysis Notations	Grade 4 (n=20)		Grade 8 (n=10)		Grade 12 (n=4)		Total (34 schools)	
	Str.	Need	Str.	Need	Str.	Need	Str.	Need
Establishing Focused Purposes	20%	64%	14%	68%	23%	63%	19%	65%
Writing for Authentic Audiences	15%	46%	11%	54%	14%	58%	14%	50%
Employing Suitable Voice/Tone	59%	3%	40%	11%	53%	9%	53%	6%
Developing Ideas	14%	58%	3%	60%	10%	59%	11%	59%
Including Supporting Details	5%	54%	1%	65%	10%	23%	5%	51%
Organizing Ideas Logically	33%	25%	32%	24%	47%	12%	35%	23%
Using Effective Transitions	4%	26%	1%	22%	2%	6%	3%	21%
Constructing Sentences	24%	9%	16%	24%	14%	5%	20%	12%
Using Language Effectively	16%	8%	7%	27%	17%	15%	14%	14%
Editing for Correctness	35%	18%	46%	9%	23%	32%	35%	18%

Areas that were noted as "Needs" at all grade levels for more than fifty percent of the participating schools were Establishing Focused Purposes, and Developing Ideas. In addition, Writing for Authentic Audiences and Including Supporting Details were "Needs" for more than fifty percent of the participating schools at two of the three grade levels. "Strength" was most frequently noted in the area of Employing Suitable Voice/Tone.

The Scoring Analysis readers also recommended which elements of the writing process they thought the school should focus on to improve future portfolio results as well as the types of portfolio entries on which schools should focus resources.

CONSISTENCY OF RECOMMENDATIONS. Double scoring and duplicate analysis of school sets was also completed to provide data concerning the consistency of recommendations for schools. Consistency of recommendations was not broken down by area e.g. Instructional Strengths and Needs, Elements of the Writing Process, and Type of Portfolio Entry. Overall, however, recommendations made by two independent groups of readers agreed 70.2% of the time for Grade 4, 72.7% for Grade 8, and 74% for grade 12. It is important to note that, without any information about the school other than the portfolios read, independent groups of teachers familiar with the writing scoring criteria made very consistent recommendations about instructional strengths and weaknesses. This signifies an important consequence of portfolio assessment -- formative evaluation of instructional practices.

For additional information concerning procedures and quality assurance measures for Scoring Analysis see the 1994-95 Writing Portfolio Scoring Analysis Final Report.

SUMMARY

Writing portfolio activities during Cycle 2 can be summarized in the form of three questions.

Is scoring accuracy improving?

Yes. The 1996 audit documents substantial improvement in local scoring accuracy. In 1993 only 23% of the audit-assigned scores agreed with school-assigned scores, and the audit average index was 40 points lower than the school average index for purposefully audited schools. Three years later, after significant training and professional development efforts, agreement between school-assigned scores and audit-assigned scores has risen to 74%, and the average index difference has decreased to 10 points for purposefully audited schools. For schools selected at random (providing a picture of overall statewide accuracy), the audit confirmed 76% of the school assigned scores with an average index difference of only 5 points on a 140 point index.

Have writing portfolio results improved?

When scoring inaccuracy is the product of inflated performance judgments, improved accuracy creates the appearance of reduced scores. Therefore the stable KIRIS results of the last three years may be interpreted as an increase in actual performance. In fact, audit data for purposefully selected schools (the sole sample available over time) demonstrate not only increases in scoring accuracy, but real increases in student performance at Grades 4 and 12. In 1993, audit results for Grade 4 schools showed an average index score of 19.6, while similar schools in 1996 showed an average index score of 46.0. Similarly, the audited Grade 12 WCI was 31.1 in 1993 and rose to 42.0 in 1996.

Can portfolios be accurately analyzed for instructional strengths and needs?

Independent readings of school samples of portfolios produced similar recommendations for future instructional emphasis and professional development. The consistency of the conclusions drawn by independent readers suggests that this analysis procedure could be very helpful to schools who want to plan effective instructional improvement based on writing portfolio contents.

SELECTED REPORTS ON WRITING PORTFOLIOS AVAILABLE FROM KDE

1993 Writing Portfolio Audit Final Report
1993-94 Writing Portfolio Scoring Analysis Report
1994-95 Writing Portfolio Scoring Analysis Report
1995-96 Writing Portfolio Audit Final Report

THE KIRIS MATHEMATICS PORTFOLIO

RATIONALE AND DESIGN. The mathematics portfolio was designed by Kentucky mathematics teachers to reflect Kentucky's instructional expectations as presented in *Transformations: Kentucky's Curriculum Framework*, and *Kentucky Content Guidelines for Mathematics*, as well as the *Curriculum and Evaluation Standards for School Mathematics* published by the National Council of Teachers of Mathematics (NCTM). The portfolio assessment differs from some aspects of traditional mathematics instruction and assessment by stressing students' choice of their best work over time with full access to resources, tools, and the evaluation criteria that will be used to judge their work.

Mathematics portfolios were implemented in 1992-93 at grades 4, 8, and 12. Starting in 1994-95 the mathematics portfolio was moved from grade 4 to grade 5 . The required contents of the mathematics portfolio include: a table of contents; a letter to the reviewer; and five to seven pieces of the student's best work. A single holistic score is assigned to the portfolio on the basis of how well the contents reflect proficiency in problem solving, reasoning, mathematical communication, and understanding and connecting the seven core concepts.

Teachers working in school scoring teams generate accountability data by analyzing and scoring portfolios using the same standard statewide. The Kentucky mathematics standard is represented by the Proficient descriptors of the Holistic Scoring Guide and associated benchmark portfolios.

PROFESSIONAL DEVELOPMENT

Since the introduction of the mathematics portfolio, KDE has supported classroom teachers through a tiered training system similar to that used in writing. A design committee composed of the regional resource teachers trains local "cluster leaders," who then deliver portfolio development and scoring training to the other teachers in their school. Two full-day workshops are produced each year. In addition, video training is provided through statewide educational television and local workshops are available from the regional consultants.

The regional mathematics consultants, supported by a grant from PRISM, work with local districts and schools upon request providing individually tailored professional development experiences that may take the form of workshops, demonstration lessons, classroom observation and feedback, portfolio analysis, or other activities designed to improve mathematics instruction and assessment.

MONITORING THE SYSTEM

No formal scoring audit has yet been conducted for the mathematics portfolio -- no portfolio review for the purpose of identifying schools with highly discrepant scores and correcting their Mathematics Cognitive Index (MCI). Scoring Analysis sessions were held in 1994 and 1995 to provide schools with informal feedback on their scoring accuracy and the contents of their mathematics portfolios. In Spring 1996, the legislature directed that the mathematics portfolio be removed from the accountability index for a period of two years to allow consideration of alternative portfolio development and scoring procedures. Mathematics portfolios prepared during the 1995-96 school year were scored locally and the results reported for Cycle 2 accountability.

SCORING CONSISTENCY

SUMMER 1994 MATHEMATICS PORTFOLIO REVIEW STUDY. This study was designed to collect information from the first years of implementation, to recommend procedures for establishing a baseline for cycle 2 (cycle 2: 1992-93 to 1995-96), and to estimate scoring accuracy. The study was limited to grades 8 and 12. This was not considered a formal audit because there was no intent to change accountability scores as a consequence of this study. Half of all Kentucky schools were invited to take part in the 1994 activity, with a similar activity scheduled for 1995 to accommodate the remainder. For each school, a set of portfolios was selected at random from the appropriate grade level by the scoring contractor. School participation was voluntary. One hundred, thirteen high schools, roughly half of all Kentucky high schools, took part. About one-third of all middle schools elected to participate (123 schools).

The procedures employed were similar to those used in the Writing Portfolio Scoring Analysis. Kentucky teachers who wished to serve as scorers had to qualify to participate in the Review Study by correctly scoring 6 of 9 portfolios (or a similar ratio equivalent to 67% accuracy). Eighty-eight of the ninety-one applicants succeeded. Table leaders were responsible for recalibrating and monitoring the scorers in their group daily. Two strategies were used to monitor scorer accuracy. First, one prescored quality control portfolio was scored each day; the table leader clarified scoring standards for any reader who was discrepant. Second, the table leader read one portfolio per scorer, per day, providing moderation as needed.

Portfolios completed by grade 8 students were available from the 1992-93 school year and from 1993-94. Grade 12 portfolios were available from the 1993-94 year only. The agreement between locally assigned scores and the Review scores was quite high, indicating that for the participating schools, local scoring accurately reflected the Kentucky standard for mathematics performance. Tables 12-6 through 12-8 show agreement rates for each group of portfolios.

Overall, 99.3% of the Review Study scores were within one level of the original scores.

STUDENT PERFORMANCE

The 1995 Portfolio Scoring Analysis Report noted that in comparison with the preceding year "scoring accuracy has remained consistent, [but] student performance has not significantly improved." Scores from all years are presented in Table 12-9. The report concluded that professional development related to core concepts was needed.

TABLE 12-6
REVIEW STUDY - ORIGINAL SCORE AGREEMENT
GRADE 8 - 1992-93 MATHEMATICS PORTFOLIOS

			Review			
Original Score	Incomplete	Novice	Apprentice	Proficient	Distinguished	Total
Incomplete	**34**	8	1	0	0	43
Novice	19	**865**	53	0	0	937
Apprentice	3	129	**77**	1	0	210
Proficient	0	8	10	**3**	0	21
Distinguished	0	1	0	0	**0**	1
Total	56	1011	141	4	0	1212
%Agreement	79.1%	92.3%	36.7%	14.3%	NA	80.8%

TABLE 12-7
REVIEW STUDY - ORIGINAL SCORE AGREEMENT
GRADE 8 - 1993-94 MATHEMATICS PORTFOLIOS

			Review			
Original Score	Incomplete	Novice	Apprentice	Proficient	Distinguished	Total
Incomplete	**5**	1	1	0	0	7
Novice	7	**706**	83	1	0	797
Apprentice	5	146	**183**	4	0	338
Proficient	0	11	44	**23**	0	78
Distinguished	0	2	3	4	**0**	9
Total	17	866	314	32	0	1229
%Agreement	71.4%	88.6%	54.1%	29.5%	NA	74.6%

TABLE 12-8
REVIEW STUDY - ORIGINAL SCORE AGREEMENT
GRADE 12 - 1993-94 MATHEMATICS PORTFOLIOS

			Review			
Original Score	Incomplete	Novice	Apprentice	Proficient	Distinguished	Total
Incomplete	**23**	3	1	0	0	27
Novice	9	**1137**	25	0	0	1171
Apprentice	2	193	**177**	6	0	378
Proficient	0	9	50	**25**	0	84
Distinguished	0	0	3	1	**2**	6
Total	34	1342	256	32	0	1666
%Agreement	85.2%	97.1%	46.8%	29.8%	33.3%	81.9%

TABLE 12-9
STATEWIDE MATHEMATICS PORTFOLIO MEAN SCORES

GRADE	1992-93	1993-94	1994-95	1995-96
8	9.0	18.1	19.5	24.8
12*	-	14.8	14.5	19.0

*For accountability purposes, 1993-94 scores were used to form the cycle 2 baseline.

USING MATHEMATICS PORTFOLIO SCORING ANALYSIS TO PROVIDE INSTRUCTIONAL FEEDBACK

1994 PORTFOLIO REVIEW STUDY. The 1994 Portfolio Review Study analyzed the presence of evidence indicating student understanding of the Kentucky Core Concepts. For the grade 8 portfolios reviewed, Mathematical Procedure was the Core Concept most developed, Mathematical Structure the least. In the grade 12 portfolios, Mathematical Procedure and Number were the Core Concepts most developed and Data the least developed. Generally, the mathematical understanding evident in the portfolios was rated as below grade level for most of the Core Concepts.

1995 MATHEMATICS PORTFOLIO SCORING ANALYSIS. In 1995, a Scoring Analysis session was held for the purpose of providing schools with feedback on their scoring accuracy and with instructional recommendations based on portfolio contents. One hundred and thirty-five grade 8 schools and 82 grade 12 schools elected to participate. Each school submitted a sample set of portfolios to be rescored and evaluated. After a school set of portfolios had been scored and discussed, the scorers identified three or four instructional strengths or needs that were reflected consistently across the portfolios and recommended which core concepts should receive added instructional emphasis.

1994 MATHEMATICS PORTFOLIO ANALYSIS PILOT. This special activity focused on the formation and training of district teams to evaluate math portfolios. Seven district teams were trained to use completed student portfolios to determine instructional strengths and areas in need of instructional change. The activity included development of district action plans for future professional development. Two different models were used. "In both models, analysis and discussion of student portfolios followed a similar introduction on standards, the scoring rubric, and the core concepts. The primary differences between the two models were the length of the training, number of participants from each district, and the opportunity to review portfolios at different grade levels." (Colopy, 1995, p. 3)

Participants were very positive about their experience. Both groups agreed that the opportunity to read and discuss portfolios across grade levels was important for understanding the logic behind the portfolio program. Follow-up interviews of participants and review of district action plans were carried out by Policy Studies Associates, Inc. The researchers note that although participants listed a number of changes that they thought must occur in order for student portfolios to improve, "participants rarely mentioned either the need to change their instructional practices to improve students' learning opportunities, or the need for professional development to improve or change teachers' instructional practices." (Colopy, p. 33)

SELECTED REPORTS ON MATHEMATICS PORTFOLIOS AVAILABLE FROM KDE

Summer 1994 Mathematics Portfolio Review Study Final Report
1994-95 Mathematics Portfolio Scoring Analysis Final Report
Colopy, K. (1995) Evaluation of the Kentucky Mathematics Portfolio Analysis Pilot. Policy Studies Associates, Inc.

THE KIRIS ALTERNATE PORTFOLIO

The Alternate Portfolio is the assessment vehicle for students with what in general terms can be described as moderate to severe disabilities that prevent them from participating in regular classroom instruction even after all possible assistance and instructional adaptations have been provided. Such circumstances likewise prevent student participation in the regular KIRIS testing program.

RATIONALE AND DESIGN

The Alternate Portfolio was developed by a team composed of teachers across the state, local administrators, and Kentucky Department of Education and University of Kentucky staff to reflect educational outcomes that are important for all students, including students with moderate to severe disabilities, and consistent with the Kentucky Education Reform Act. Eligibility for the Alternate Portfolio is based upon the student's need for an individualized curriculum based upon 'life-skill' needs.

The Alternate Portfolio includes a Table of Contents, a Letter to the Reviewer from the student which may be written with the help of his/her teacher and non-disabled peers, a letter from the parent with their thoughts on the student's portfolio, and 7 to 10 entries. Portfolio entries include the student's school schedule written with photographs or pictures so that the student can understand it, examples of how the student communicates with others, a school and community job resume for the student if he/she is listed as an 8th or 12th grader, and other entries selected and developed by the student.

The Alternate Portfolio is scored at grades 4, 8, and 12. For a student enrolled in an ungraded program, the portfolio is scored in the year that the student is nine years old as of October 1, 13 years old as of October 1, or during his/her last year in school.

The Alternate Portfolio employed a three-level moderation scoring system. Each student's teacher scored the portfolio (district), and then the portfolio was scored a second time by another teacher from within the same region of the state. If they did not agree and the original teacher-assigned score was Proficient or Distinguished, the portfolio was rescored again at the state level to establish a final score. Any time there was a difference between district and regional scoring, a district was permitted to request that the portfolio be rescored at the state level. If the scores which were assigned during the first two rounds of scoring for any Alternate Portfolio differed by two or more performance levels, the portfolio was rescored at the state level.

MONITORING THE SYSTEM

Scoring consistency is based on agreement between the original, district-assigned score and the regional score assigned by a teacher who was generally from a different district. The scorers of Alternate Portfolios receive training intended to be similar to that used for the writing portfolio. However, the number of scorers of Alternate Portfolios is much smaller than that needed for writing or mathematics portfolio. Writing portfolios, for example, are administered to approximately 140,000 students annually as compared to fewer than 1,000 Alternate Portfolios annually.

During Cycle 2, scoring accuracy could also have been viewed as agreement between district and regional scores in that the regional scorers had undergone the same training but were not personally involved with the students developing the Alternate Portfolio as was the person doing district scoring -- at least, in most cases. The agreement between either district or regional scores with scores assigned at the state level are not useful as an accuracy index because not all portfolios were rescored at the state level and the selection of portfolios that was rescored at the state level was not a random process.

Table 12-10 summarizes scoring consistency and accuracy based on comparison of district and regional scores. contains summary information from these tables.

Both scoring consistency and accuracy declined for the Alternate Portfolio from 1993 to 1996. Exact agreement between local scores and regional scores dropped by 11%. During the same period, local scoring appears to have become less accurate for all grades assessed. Of the local scores changed by regional scorers in 1993, slightly more than half were reduced. By 1996, the number of scores reduced by regional scorers was roughly four times the number raised. Across all grades, the reduction amounted to 18 points on the 140-point KIRIS scale. The cause of the pattern is not clear. Scoring training and materials did not changed substantially during this period.

More important than the differences in the delivery of training may be the difference in the opportunities a trained teacher is likely to have to apply the training to real portfolios. Most teachers trained to score the Alternate Portfolio will encounter only 1 to 5 portfolios in a year, and probably will not score alternate portfolios for accountability purposes every year. Observers report that teachers of students qualifying for the Alternate Portfolio tend not to focus on the portfolio development (or at least the scoring) until the accountability year. That may mean that the standards were not well understood by the teacher until the accountability year and/or that entries prepared for the portfolio during non-accountability years did not reflect the standards.

REPORT ON ALTERNATE PORTFOLIOS AVAILABLE FROM KDE

Alternate Portfolio: Scoring Consistency, Accuracy, and Recommended Changes

TABLE 12-10
SUMMARY OF AGREEMENT BETWEEN DISTRICT AND REGIONAL SCORING OF ALTERNATE PORTFOLIOS

	GRADES			
	All	4	8	12
YEAR	PERCENT EXACT AGREEMENT			
1993	63.42%	62.98%	67.96%	56.61%
1994	50.28%	55.59%	49.28%	44.81%
1995	52.05%	54.60%	52.04%	48.33%
1996	52.26%	50.00%	53.30%	53.88%
	% SCORES RAISED BY REGION			
1993	16.71%	21.37%	14.24%	14.29%
1994	12.51%	12.78%	11.75%	13.28%
1995	10.64%	11.66%	9.40%	11.00%
1996	9.48%	10.61%	8.97%	8.68%
	% SCORES LOWERED BY REGION			
1993	19.87%	15.65%	17.80%	29.10%
1994	37.21%	31.63%	38.97%	41.91%
1995	37.31%	33.74%	38.56%	40.67%
1996	38.25%	39.39%	37.73%	37.44%
	RATIO RAISED/LOWERED SCORES			
1993	0.84	1.37	0.80	0.49
1994	0.34	0.40	0.30	0.32
1995	0.29	0.35	0.24	0.27
1996	0.25	0.27	0.24	0.23
	DISTRICT ASSIGNED INDEX			
1993	22.7	18.9	20.1	32.1
1994	59.0	52.1	57.3	70.3
1995	66.8	66.3	61.6	75.1
1996	65.8	68.0	67.6	67.6
	REGIONALLY ASSIGNED INDEX			
1993	21.1	21.8	18.5	24.3
1994	44.7	41.7	42.6	51.6
1995	50.6	52.6	44.4	56.8
1996	47.5	48.7	48.6	48.6

PORTFOLIO ISSUES SUMMARY

Despite the differences in scoring activities and portfolio types, the data present an interesting pattern in scoring accuracy across all three portfolios which might be summarized as "Practice makes perfect." Teachers in the accountability grades score many writing or mathematics portfolios each year and their agreement with external scorers is generally greater than 70% and improving over time. Moreover, they are faced every year with the challenge of shaping their instruction to align with portfolio scoring criteria. The Alternate Portfolio, which is infrequently encountered, has been less robust in terms of scoring accuracy.

The same effect is evident within the writing or mathematics scoring data. Locally-assigned scores are consistently most accurate at the performance levels that occur most frequently. As one observer noted. "After you have scored 300 Novice portfolios, you have a pretty good idea of where Novice begins and ends. And after reading 300 Apprentice portfolios, you have a pretty good idea of what they look like. But probably no one [local scoring team] in the state has yet scored 300 Proficient portfolios. It's really not surprising that accuracy is a little lower there." The implications for ongoing scoring training are clear. Portfolio scorers need many opportunities to read and score portfolios, particularly portfolios at performance levels beyond those encountered in their school or classroom. Despite the difficulties of locating large numbers of Proficient and Distinguished portfolios, KDE must find and distribute more examples of student work at these performance levels.

The connections between portfolio assessment, related professional development, and instructional practice have not yet been well documented. The fact that independent groups of trained readers draw the same conclusions about instructional program strengths and weakness after scoring a school set of portfolios is important. It shows one way for a school to use portfolio analysis information to plan relevant professional development and instruction for the coming year. When a standard procedure for portfolio analysis is used across schools, the state can begin to link instructional practices with improved portfolio scores. KDE should build on the results from the Mathematics Analysis Pilot to find the best model for training district teams to carry out an extended portfolio scoring process that includes program evaluation and instructional planning.

The Colopy study points out that until teachers acknowledge that instructional change is the key to improving student performance, one of the major benefits of portfolio assessment will be unrealized and results unlikely to change. On the other hand, high gains in writing portfolio scores from several schools were validated by the 1996 audit. Many of these schools were smaller than average. Speculation about this fact was answered by teachers' explanation that small schools find it easier to adopt an intense, schoolwide focus on a particular program, and it is this planned alignment between assessment criteria, professional development, and instructional change that accounts for large portfolio score gains. Clearly, these schools used portfolio resources to full advantage, but this practice is not yet widespread. KDE efforts to document the portfolio practices of high performing schools and to build that information into statewide portfolio training activities must continue so that all schools can effectively use portfolio assessment to document and to improve student performance.

SECTION IV
GENERALIZABILITY

Section 4 examines issues related to test score validity in a broad sense. Chapter 13 presents reliability estimates for several components of the test. Chapter 14 describes existing validity-related evidence for KIRIS. Chapter 15 explores consequential validity for the assessment, including differences in performance at the student level by gender and ethnicity.

CHAPTER 13
RELIABILITY

STUDENT-LEVEL RELIABILITY

Although accountability decisions take place only at the school level, results are reported to parents at the student level. It is useful, therefore, to examine reliabilities at that level. Table 13-1 lists student-level coefficient alphas for open-response items for the second accountability cycle. Alpha was computed by form for common and matrix items combined. The median value of alpha as well as the range was then assessed, based upon values obtained across the 12 forms. These figures were calculated using results from students who were eligible to complete open-response tests and who were present on the day of testing, with absence defined as blank responses to all common and matrix items; blank responses by students who had attempted at least one question were counted as zero in the computation of coefficient alpha. The reason for not including the responses of absent students is that score reliability would be slightly over-estimated.

Although the Department of Education advises against making student-level decisions based on individual test scores alone, open-response test reliabilities compare favorably with reliabilities from other tests used to make student-level decisions. Composite score reliability is comparable to reliabilities noted for the ACT Composite, and individual subject area reliabilities are similar to ACT and CTBS subject area reliabilities.

As a caveat, it should be noted on the one hand that using coefficient alpha probably underestimates score reliability insofar as item raw scores are the basis for the computation. A seven or eight item test having a single, relatively difficult item is likely to yield a somewhat diminished coefficient alpha in comparison with results from a comparable test having essentially uniform, moderate item difficulty. By contrast, the fundamental scaling method used by KIRIS employs a logistic model. The use of item response theory takes into account differences in item difficulty not reflected in the computational use of raw scores.

On the other hand, coefficient alpha does not take into account variability in student-level scores arising from the fact that not all students have the same scorer. As indicated in Chapter 8, the effect of having different scorers appears minimal even at the student level, with correlations among ratings assigned by two scorers to the same student's responses in a given content area ranging from .87 to .99. Combined with the fact that scoring is monitored through a quality assurance process, these results indicate that the effect of scorers on a given student's scores is very small, and probably negligible for the KIRIS composite. At the same time, use of one scorer for all 7 or 8 of a student's responses in a given content area initially suggests the possibility that coefficient alpha could be inflated by a within-student scorer effect. To test this possibility, coefficient alpha was recomputed using the sample of grade 11, 1995-96, students whose papers were scored twice. Median coefficient alpha, across forms, for scores to odd-numbered items assigned by the first scorer and scores to the even numbered items assigned by the second scorer, was highly similar to results in Table 13-1; two content areas yielded slightly lower alpha, two others yielded slightly higher alpha, and composite score alpha was unchanged. Thus, scorer effects do not appear to inflate coefficient alpha estimates systematically.

Because arts & humanities and practical living/vocational studies items are used for purposes of school-level score reporting rather than student-level score reporting, no estimates of student-level score reliability are provided. Coefficient alpha cannot be computed for writing and mathematics portfolio scores at all, given that each portfolio receives only one score, and at least two scores per student are necessary to derive a student-level reliability statistic. See Chapter 12 for portfolio scoring consistency data.

TABLE 13-1
OPEN-RESPONSE TEST RELIABILITIES BY SUBJECT AND GRADE

Grade	Subject	1992-93* Coefficient Alpha		1993-94* Coefficient Alpha		1994-95* Coefficient Alpha		1995-96* Coefficient Alpha	
		Med.	Range	Med.	Range	Med.	Range	Med.	Range
4	Reading	**.82**	.81 - .83	**.82**	.81 - .84	**.81**	.80 - .85	**.81**	.79 - .83
	Mathematics	**.78**	.75 - .80	**.80**	.77 - .83	**.82**	.81 - .84	**.81**	.80 - .83
	Science	**.72**	.71 - .75	**.78**	.76 - .80	**.76**	.72 - .78	**.75**	.73 - .77
	Social Studies	**.76**	.72 - .77	**.80**	.78 - .84	**.80**	.78 - .82	**.83**	.80 - .84
	Composite	**.92**	.91 - .92	**.93**	.93 - .94	**.93**	.93 - .94	**.93**	.93 - .94
8	Reading	**.85**	.81 - .87	**.84**	.82 - .86	**.85**	.82 - .86	**.85**	.84 - .86
	Mathematics	**.82**	.78 - .83	**.78**	.76 - .80	**.84**	.81 - .86	**.85**	.82 - .86
	Science	**.78**	.74 - .80	**.74**	.70 - .76	**.81**	.77 - .82	**.80**	.78 - .82
	Social Studies	**.83**	.81 - .87	**.86**	.80 - .88	**.88**	.86 - .89	**.86**	.84 - .88
	Composite	**.94**	.93 - .94	**.93**	.93 - .94	**.94**	.94 - .95	**.95**	.94 - .95
11/12	Reading	**.84**	.81 - .87	**.82**	.78 - .83	**.85**	.84 - .87	**.85**	.83 - .86
	Mathematics	**.82**	.80 - .85	**.80**	.78 - .85	**.87**	.85 - .88	**.84**	.83 - .85
	Science	**.83**	.80 - .84	**.81**	.76 - .83	**.84**	.80 - .87	**.82**	.80 - .84
	Social Studies	**.86**	.83 - .87	**.88**	.85 - .88	**.87**	.87 - .89	**.85**	.83 - .87
	Composite	**.94**	.94 - .95	**.94**	.94 - .95	**.96**	.95 - .96	**.95**	.94 - .95

* Median coefficient alpha based upon common and matrix items across 12 forms.

SCHOOL-LEVEL RELIABILITY

Like student-level reliability, the main concern for school-level reliability is *consistency* of results. However, consistency of results with respect to schools is more complex in that consistency centers around the accountability decisions made about a school. More specifically, to what extent do the results obtained in a given year on the basis of the particular test items used that year and the particular students within each school in an accountability grade generalize to a theoretical pool of other tasks that might have been selected and to other students who might have attended an accountability grade? Put another way, had an alternative set of questions from the same test domains been administered to an alternative set of students who might have attended the school, to what extent would the school have obtained the same accountability rating for that particular accountability cycle. This type of *decision* *consistency* is the most important issue relative to school-level reliability.

The KIRIS accountability system requires the computation of an accountability index for each accountable school and district in the state. The accountability index is the weighted average of eight indices. Seven of the indices indicate performance in cognitive areas (reading, writing, mathematics, science, social studies, arts & humanities and practical living/vocational studies) derived from results of open-response testing and/or portfolios. The eighth index is a combination of noncognitive factors used to measure the effectiveness of a school. Because schools are judged by their improvement over time rather than their status at any one time, the school accountability statistic, which is based upon the accountability index and is the basis for reward and assistance decisions, is a gain score. Gain scores are calculated for a four year accountability cycle in which a baseline is set by averaging the first two years worth of data, an improvement goal is established, and the average of the two succeeding years of data are compared.

Because it was neither reasonable nor feasible to make students take KIRIS twice in each year of the second accountability cycle, no data is available to estimate accountability index gain score reliabilities using test-retest calculations. To estimate the amount of error in accountability index gain scores, it was necessary to conduct a simulation study. The strategy of the study described below was to estimate variability in scores due to measurement error and sampling error in the first stage, using schools sufficiently large to support a statistical technique known as sampling with replacement. This first stage yielded an estimate of gain score standard error. The second stage used actual school score and size data to simulate variability in scores for each year of the accountability cycle one hundred times, comparing the school accountability classification obtained in each of the hundred replications with the classification assigned in reality. The second stage yielded an estimate of decision consistency.

ACCOUNTABILITY INDEX GAIN SCORE STANDARD ERROR OF MEASUREMENT. To estimate the standard error of measurement for the KIRIS accountability index gain score, a computer simulation was performed using actual school results for Accountability Cycle 2. To simulate the variability in obtained scores associated with the students enrolled in an accountability grade in a school in a particular year relative to those who could attend if they had been a little younger or a little older, multiple sets of students were obtained by drawing samples from each school with replacement. To simulate the variability in obtained scores associated with the items that appear on KIRIS tests each year relative to the items that could appear, each time a student was selected to be in the simulation study, the student's observed score was resampled, or changed, to simulate the fact that the student would probably score somewhat differently on a different set of test items. (For portfolio scores, this resampling of student score was accomplished by sampling from the possible scores that might have been obtained if a different teacher had scored the portfolio.) This first kind of variability will be referred to as sampling error (or student sampling error), whereas the second kind will be referred to as measurement error. Note that variability in noncognitive indicators was simulated as well.

Using schools sufficiently large to provide for meaningful resampling of students, accountability index scores were computed for each school 50 times. To estimate sampling error and measurement error, pooled within-school variance estimates were computed across the 50 scores after the first step of the simulation (drawing multiple sets of students) and the second step (resampling or changing individual test scores for sampled students). As a check on the process, the second step was also conducted without the first step, i.e., individual test scores were changed 50 times for the same set of sampled students. This procedure yielded an estimate of the gain score standard error of measurement expected for each school size included in the study, i.e., 24, 36, 48, and 96. (Multiples of 12 were selected to accommodate student resampling within form, which was necessary to reflect the matrix-sampled design used in KIRIS.)

Gain score standard error of measurement figures for schools of size 96 in all four years of Accountability Cycle 2 are presented in Table 13-2. Larger gain score standard errors were found for smaller sample sizes, and were almost exactly proportional to the theoretically expected value, based on the inverse of the square of sample size. Finding a standard error of measurement due to sampling error substantially greater than standard error due to measurement error, i.e., twice as large a standard error, is consistent with previous results (see KIRIS Accountability Cycle I Technical Manual).

TABLE 13-2
ESTIMATED ACCOUNTABILITY INDEX GAIN SCORE STANDARD ERROR OF MEASUREMENT FOR SCHOOLS OF N = 96 STUDENTS

	Sources of Score Variability Considered		
School Level	Sampling Error	Measurement Error	Measurement and Sampling Error*
Elementary	1.49	.76	1.42 (1.67)
Middle	1.61	.71	1.46 (1.76)
High	1.52	.71	1.43 (1.68)

*The first figure given in this column provides the actual results obtained in the study in a two-step simulation of sampling error followed by measurement error, whereas values in parentheses show the result obtained through additive combination of separately obtained sampling and measurement error variance estimates.

As suggested by the footnote to Table 13-2, an unusual result was obtained in the study when pooled within-school variance results were computed from school scores obtained after applying both procedures, i.e., sampling of students (sampling error) and changing student scores (measurement error). That is, a lower standard error of measurement was obtained across school scores incorporating measurement error and sampling error, than across scores incorporating sampling error alone. This result is somewhat paradoxical, although it may be a legitimate outcome of using a performance level based scheme approach to the particular data in Accountability Cycle 2. That is, replications of schools yielding an unusual draw of students, such as a very strong or very weak group of students relative to the population, may yield scores that are drawn inward through the simulation of measurement error. Student scores assigned to the Novice performance level can only obtain a performance level higher than or equal to Novice in the simulation, and those assigned Distinguished can only obtain a performance level lower than or equal to Distinguished. In any case, a more conservative result (one suggesting a greater amount of error) was obtained through additive combination of variances associated with the standard error due to student sampling error estimated in the first step, and the independently estimated standard error due to measurement error.

An essential description of the procedures followed in the study is contained in "Effects of Students and Tasks on Gain Scores Used in Complex School Accountability Decisions" (Dings & Kingston, 1996). An updated and more detailed description of the study will be issued by KDE.

ACCOUNTABILITY DECISION CONSISTENCY. Having obtained standard error of measurement estimates of the likely magnitude of error in accountability index gain scores for schools of the sizes simulated in the study, which were held at the same size over the course of the four year accountability cycle, assessing decision consistency for schools that vary in the number of students at the accountability grade over the four years required an application of standard error of measurement results. To estimate decision consistency in a manner that takes into account variability in the number of students enrolled at the accountability grade each year, it was necessary to consider the obtained accountability index score in each year, accountability grade enrollment, and an estimate of error for the accountability index. The first two pieces of information are a matter of record, whereas the latter was obtained through application of results presented above.

Applying the above results was relatively straightforward, to the extent that it is reasonable to assume standard error of measurement for a single year accountability index score is equal to the standard error of measurement for a gain score computed by subtracting the average of two single year indices from the average of two other single year indices. All other things being equal, doubling the number of years from one to two reduces standard error of measurement by a factor of the square root of 2, but relying on a gain score increases standard error of measurement by the same factor. In this regard, it was reasonable to use figures presented in Table 13-2 as standard errors of measurement for a single year accountability index when 96 students are present in the accountability grade. Error estimates for other accountability grade enrollments are obtained through multiplication by the square root of 96 and division by the square root of the actual enrollment.

In estimating decision consistency, accountability index scores were computed for each of the four years of Accountability Cycle 2 as draws from a normal distribution with mean equal to the school accountability index mean for a given year and standard deviation equal to the standard error of measurement associated with accountability grade enrollment in that year. Results from each year were averaged appropriately (weighted by accountability grade enrollment) to obtain a baseline, a growth index, and a corresponding accountability decision for each of one hundred replications. Decision consistency was assessed by comparing results of the one hundred replications with the original accountability classification assigned to each school.

Given that some degree of uncertainty is inherent in any measurement, it is apparent that perfectly consistent accountability decisions are not possible. However, the consequences of a decision typically influence how consistent we need to be in making that decision, i.e., how much uncertainty or inconsistency we can tolerate. KIRIS results are used to assign cash rewards to schools whose student test scores exceed improvement goals and to assign the sanction of "school-in-crisis" to schools whose scores decline substantially. To protect the investment of taxpayer dollars, we want to be highly confident that a school's achievement has truly improved if it is to be rewarded. Similarly, to protect the opportunity for local control of schools and to avoid the negative stigma of sanctions, we want to be highly confident that a school's achievement has truly declined if it is to be designated a "school-in-crisis." (The reasoning employed in this regard is similar to that which applies to the College Board's Advanced Placement (AP) Examinations, used by many colleges to award course credit if a student scores 3 or higher on a 0-5 point scale. Because a score of 3 is the decision point, the college granting credit wants to be reasonably confident that a student who obtained a score of 3 would not have received a score of 2 or lower on retesting.)

Based upon concern for the expenditure of taxpayer dollars and the preservation of opportunity for local control of schools, the following two questions, and accompanying results, are important relative to Kentucky's accountability program:

1. What percentage of schools categorized as "Reward" truly improved their achievement?

   Based on resampling analyses for accountability cycle two, over 99 percent of the time, schools in the simulation originally assigned to the "Reward" classification obtained scores in the simulation in which the growth index had increased relative to accountability baseline. About 94 percent obtained scores in which the growth index had met or exceeded the improvement goal derived from the accountability baseline.

2. What percentage of schools categorized as "In Crisis" truly did not improve their achievement?

   Over 99 percent of the time, schools in the simulation originally assigned to "In Crisis" obtained scores in the simulation in which the growth index had NOT increased relative to or equaled the accountability baseline.

The full set of decision consistency results is given in Table 13-3, in which conditional percentages are presented. Figures in each row represent the proportion of schools originally classified as the heading in that row for which the accountability decision noted in the column was obtained. The second entry in the top row is correctly read as meaning .134 of replications of schools originally classified as being in crisis yielded an accountability classification of Decline in the simulation study. These results indicate an acceptable level of decision consistency, particularly with regard to the misclassifications (specified above) that are most important to avoid. The "Success" classification is the only inconsistently replicated accountability decision, a result which is expected based on the fact scores leading to a classification of "Success" encompass a very short range of the accountability gain score scale: a single point, from just above "Improving" to just below "Rewards". The very notion of the "Success" classification is that of a buffer between receiving a very modest sanction (having to write a school improvement plan) and a reward. The remaining classifications yield consistent decisions, with the two bounded categories ("Improving" and "Decline") obtaining results identical to their original decisions .81 of the time, and the unbounded categories ("In Crisis" and "Rewards") yielding such results .87 of the time.

It should be noted that decision consistency results are based on decisions for elementary, middle and high schools taken separately, and do not include decisions based on combined elementary/middle or middle/high grades. Also, in keeping with KDE's position that school scores based on extremely small numbers of students are too greatly subject to sampling error to be used with the desired level of confidence for purposes of school accountability, classification consistency is based only on schools of size twenty or larger in each of the four years of the second accountability cycle. The school size of twenty was chosen somewhat arbitrarily for purposes of implementing this study as a round number in the general vicinity of the point at which the desired level of confidence probably is not obtained, and should not be confused with or taken as an indication of KDE policy. There is, generally speaking, less uncertainty about decisions made regarding large schools, and about schools with exceptionally high score gains or large score declines. Standard error of measurement information may be used to quantify, based on school enrollment, the degree of uncertainty of a particular accountability decision relative to possible other decisions.

TABLE 13-3
CONDITIONAL PROBABILITIES ILLUSTRATING ACCOUNTABILITY DECISION CONSISTENCY

Original Decision	Proportion of Original Decisions In Which Each Accountability Decision Occurred in Simulation				
	In Crisis	Decline	Improving	Success	Reward
In Crisis	.866	.134	.000	.000	.000
Decline	.038	.813	.148	.000	.000
Improving	.000	.058	.813	.069	.061
Success	.000	.000	.412	.239	.349
Reward	.000	.000	.065	.069	.867

Note: Due to rounding, row percentages do not sum to 1.000 in every case. Column percentages are not intended to sum to 1.000, as the table provides conditional probabilities within each row.

CHAPTER 14
VALIDITY-RELATED EVIDENCE

In a statement endorsed by national associations of psychologists, researchers and practitioners of education and educational measurement, the Standards for Educational and Psychological Testing[1] describes validity as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores." The Standards also describes three broad, traditional categories of evidence necessary to support such inferences: content-related, criterion-related and construct-related evidence of validity. Evidence of these kinds for the use of KIRIS in evaluating student performance at the school level is presented below.

In the twelve years since the Standards, traditional notions of validity have been supplemented with specific criteria for performance assessments (e.g., Frederiksen & Collins[2], Linn, Baker & Dunbar[3]) as well as the idea that the consequences of using a given test are an important aspect of validity. Validity addresses the issue of whether a test measures what it is intended to measure, or even more broadly, whether it has the effect it is intended to have. In a high-stakes accountability system such as KIRIS, it is that larger definition that is of paramount importance. This type of validity, termed "consequential validity," was advocated by Samuel Messick[4] of Educational Testing Service, who is greatly responsible for the wide use of the term.

ISSUES REGARDING CONSEQUENTIAL VALIDITY

The importance of the consequential validity of the test becomes obvious when one thinks of the alternatives. KIRIS exists because the Kentucky legislature believed that KIRIS would increase the likelihood of KERA's success by supporting (indeed, driving) changes in the classrooms of Kentucky. Thus, all decisions related to KIRIS ultimately must be considered in light of the question, "Will this change lead to better instruction and more real achievement?" If not, the change becomes less justified.

Traditional thinking about reliability and validity could lead to an assessment design contrary to the true purpose of KIRIS: supporting real change in the classroom. If the tests were highly reliable and valid in the traditional sense, but higher test scores reflected only superficial changes motivated solely by the goal of raising those test scores, KIRIS would not be fulfilling its obligation. Of course, if the tests could not demonstrate sufficiently high levels of traditional reliability and validity, people would probably not view them as fair, and therefore it is unlikely that the tests would have the desired impact on classroom practice. Therefore, high degrees of reliability and validity in the traditional sense are probably **necessary** conditions for high consequential validity, but they certainly are not **sufficient** conditions. That is, high consequential validity is probably not possible without high degrees of traditional reliability and validity, but high degrees of traditional reliability and validity cannot guarantee high consequential validity.

---

[1] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington: APA. Note that a revised edition of the Standards is in preparation.

[2] Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27-32.

[3] Linn, R. L., Baker, E. L. & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.

[4] Messick, S. (1989). "Validity", in R.L. Linn (ed.), *Educational Measurement, Third Edition*, MacMillan Publishing Co., 1989.

Eventually, the success of KIRIS will be judged on the extent to which classroom practices have improved throughout Kentucky. If teachers merely make modest changes in their teaching in order to raise test scores, and increases in KIRIS scores are not reflected in other measures of learning, KIRIS will not be judged successful no matter what reliability or validity coefficients it might be able to demonstrate. However, that judgment is far in the future; thus far, KIRIS has only made a small portion of its potential impact.

THE ROLE OF RELIABILITY IN SELECTING KIRIS ASSESSMENTS

The designers of KIRIS attempted to balance the need for reliability with consequential validity. Deciding whether to include a given type of assessment was guided by the following principle: The overall Accountability Index must have sufficient reliability, but once that criterion has been met, the decision of whether to include an assessment, and if so, how much to weight it, was based on the assessment's presumed impact on the change in teaching.

This is one reason why so much importance is placed on improving the reliability of the portfolios in the assessment system. In traditional forms of measurement, classroom teachers are passive observers of the testing program. With portfolios, they are active participants. Teachers must understand fully both the content standards and the performance standards (what is to be learned and how well it must be learned) if their students are to develop quality portfolios and the teachers are to evaluate them. Portfolio reliability should increase over time to the point where their place in the assessment system is unquestioned.

TRADITIONAL VALIDITY-RELATED EVIDENCE, WITH NOVEL CRITERIA

Although consequential validity concerns may ultimately prove more important than issues of technical quality, it remains very important to examine KIRIS score validity from a traditional psychometric perspective. Thus, content validity information is reviewed below, and construct-related validity evidence is discussed on the basis of the relationships of KIRIS tests with each other, with scores from other testing programs, and with qualitative criteria for judging school quality. Intermingled with traditional notions of validity in this analysis are more recently proposed criteria for evaluating performance assessments: systematic validity, directness and transparency[5]; and, fairness, transfer, generalizability, cognitive complexity, content quality, content coverage, meaningfulness, cost, efficiency, and consequences[6].

CONTENT-RELATED VALIDITY EVIDENCE

Section II of this manual describes how the components of the KIRIS assessment were derived from Kentucky learner outcomes and academic expectations, using advisory committees of Kentucky teachers to make those outcomes and expectations operational through test items, and to make choices about what the tests would contain. Several tables summarize the learner outcomes and academic expectations measured by the items in the assessment, and there is little to add to the extensive treatment of this material. In short, there is substantive validity-related evidence in the process by which KIRIS assessments were constructed.

---

[5] Frederiksen & Collins.

[6] Linn, Baker & Dunbar.

While test development information serves as the primary source of content-related validity evidence, examining KIRIS tests in terms of the novel content-relevant criteria noted above provides a potential source of additional evidence. Cognitive complexity, content quality, and content coverage can serve as criteria by which to evaluate performance assessments. Because there exist no established standards for these criteria (as noted by Linn, Baker and Dunbar), any detailed consideration of them probably requires discourse substantiated by expert judgment in the form of task analysis. As noted in the KIRIS Accountability Cycle I Technical Manual, content coverage by open-response items in KIRIS was found to be acceptable as judged against the coverage afforded by multiple-choice items for the particular number of each type of question administered in 1993-94. Although content quality *per se* has not been examined, the presence of teacher and other content area specialist judgment in writing and selecting items for the assessments provides some indication of validity in this regard. Arguably, most performance events embody essential characteristics of cognitively complex tasks, requiring the student to demonstrate problem-solving.

CONSTRUCT-RELATED VALIDITY EVIDENCE

Because the KIRIS testing program assesses student performance in several content areas using a variety of testing methods, it is important to study the pattern of relationships among such content areas and testing methods. One method for studying patterns of relationships to provide evidence supporting the inferences made from test scores is the multitrait, multimethod matrix (see KIRIS Accountability Cycle I Technical Manual). Another method for studying patterns of relationships among varying types of test or item scores is factor analysis.

To provide evidence for the construct validity of KIRIS open-response item scores, factor analysis was performed on data obtained from open-response scores from the 1992-93 through 1995-96 KIRIS testing program. These analyses employed exploratory methods because this was an initial exploration of the factor structure of item level data from a complex, multifaceted assessment system. Responses of students at: grades 4, 8 and 11/12 to questions in the content areas reading, mathematics, science and social studies were included in the analyses. Because multiple choice items were not included in the calculation of Cycle 2 accountability indexes, and yielded student-level scores for only the 1992-93 and 1993-94 school years, they were excluded from the factor analyses. In addition, performance events were not included due to the matrix-sampling approach used in the testing program, whereby each student (in nearly every case) completed a single performance event. (Logistical difficulties associated with matching student-level performance event data and open-response records would have further limited consideration of performance event data.) Matrix-sampling of varied numbers of arts and humanities and practical living/vocational studies items in the 1992-93 and 1993-94 test administrations, for which these content areas were assessed using "integrated" items, also posed difficulties that led to their not being included in factor analytic studies. Finally, because the main focus of the factor analysis was to look for evidence of subject specific achievement factors and general factors at the *item* level, holistic scores from the writing and mathematics portfolios were not considered. (Note that correlations involving student-level portfolios multiple-choice item scores were examined in the KIRIS Accountability Cycle I Technical Manual, as were school-level performance event correlations.)

The two main research questions addressed below are: (a) what are the relationships among KIRIS content areas as measured by open-response items and (b) are such relationships consistent across KIRIS accountability grades and years. While the present analyses represent an initial exploration of item level data, there are several general hypotheses about the factor structure of the data, especially with regard to the research questions and what observed factor structure would count as favorable or unfavorable evidence relative to item score validity. Before addressing these underlying hypotheses and the specific methods used in the present research, it is more appropriate to first consider in more detail the open-response questions used in the KIRIS testing program.

The open-response questions used in KIRIS require students to compose or generate their answers rather than respond simply by recall or recognition. In line with KERA's educational direction, "they emphasize student communication skills, and thus encourage teachers to use more powerful instructional practices that utilize communication skills as well as other aspects of problem solving" (KIRIS Accountability Cycle I Technical Manual, p. 45). Items were designed to be completed by students in approximately ten minutes.

As noted above, open-response items were administered in reading, mathematics, science and social studies. In each content area, each student was administered 5 to 6 common items (i.e., items administered to all students) and 2 matrix-sampled items (i.e., items unique to a particular form of the test). Because the factor analysis procedures used require complete responses from all examinees, matrix-sampled items were excluded from the analyses. In summary, the present research studied the relationship among open-response, common items in reading, mathematics, science, and social studies.

One hypothesis relating to the factor structure of the above data is that four method factors will be found such that each factor has items loading from one and only one content area. Of course, this hypothesis would not be well supported by the emergence of a single factor in which all open-response items loaded, or, more particularly, by the emergence of multiple factors with loadings not corresponding to the individual content areas. Conversely, the finding of content factors such as a mathematics content factor and/or a reading content factor would be considered favorable. Traditionally, high correlations among scores in the same content area has been considered evidence supportive of correlation-based validity.

Assuming relationships are found among KIRIS content areas, a second research question is of interest regarding the consistency of those relationships across KIRIS accountability grades (grades 4, 8 and 11/12) and years. One hypothesis is simply that the factor structure observed will be the same across grades and years. However, this hypothesis may be too simple. For example, there can be little doubt that instructional practices differ across grades. With respect to the KIRIS accountability grades, instructional practices at the elementary level (grade 4) are different than at the middle (grade 8) and secondary (grades 11/12) levels. Therefore, a second hypothesis regarding the factor structures found across KIRIS accountability grades is that such factor structures will vary across grades. The lower the grade level, the more likely it is for students to have the same teacher providing instruction across content areas. This particular difference in teaching practices across grades suggests that the lower the grade, the greater the influence of method. As such, if factor structures are found to vary across grades, it is hypothesized that at earlier grade levels a single factor will emerge in which open-response items across content areas load.

METHOD. A factor analysis was performed on the open-response, common item scores obtained from each year and grade of the KIRIS testing program for Cycle 2. Students from each of the KIRIS accountability grades obtaining a complete set of open-response, common item scores (i.e., no missing scores) were included in the analyses and all analyses were performed using SAS (SAS Institute Inc., 1989a, 1989b, 1989c). In the factor analyses discussed below, a non-iterative, principal factors approach was used as the initial factor extraction technique. An iterative approach was not used in the present analyses for reasons of economy. However, an iterative approach performed at one grade level showed results to be very similar to the non-iterative approach. Squared-multiple correlations were used as prior estimates of communality and an oblique rotation to the promax criterion was used to rotate each of the initial factor solutions. The interpretability of the rotated factor solutions served as the primary criterion for selecting the final rotated factor pattern matrices to discuss.

RESULTS. Because the factor analytic results found across KIRIS grades and years were so striking similar, the hypothesis that the observed factor structure would be the same across grades and years was confirmed. As such, with regard to the second research question, there appears to a considerable degree of consistency across KIRIS accountability grades and years. For this reason, Tables 14-1 and 14-2 present the results from only one grade and year, i.e., grade 8, 1996. The factor analytic results for grade 8, 1996 are very representative of the results obtained for grades 4 and 11/12, and for all years of Accountability Cycle 2. The analysis presented in Table 14-1 is based upon 50,321 students. The inter-factor correlations presented in Table 14-2 are based upon the same group of 8th graders as the results presented in Table 14-1.

Equally as striking as the degree of consistency across KIRIS accountability grades and years was the degree of simple structure obtained in the four-factor rotated solutions. (Inspection of the five-factor rotated solutions for each grade level and year revealed no substantial loadings on the fifth factor.) Note that the rotated pattern loadings in Table 14-1 can be interpreted as regression coefficients or weights and thus represent the effect of a factor on an item score with the effects of the other factors partialed out (Rummel, 1970). For grade 8, 1996, factor loadings above .40 have been highlighted (bold type) within Table 14-1. Inspection of Table 14-1 reveals that each rotated factor is clearly defined by loadings on one and only one KIRIS test. For example, factor I is defined by the factor loadings for reading and factor II is defined by the factor loadings for mathematics. Factors III and IV are similarly defined for social studies and science, respectively. Note also that the factor loadings in Table 14-1 do not appear to be artifacts of scoring procedure; comparable results are obtained when this analysis is replicated using the sample of grade 8, 1995-96 students whose papers were scored twice, based on scores assigned to the odd-numbered items by one scorer and scores assigned to the even numbered items by a different scorer.

Although the rotated four-factor solutions seem very interpretable, it also appears that a general achievement factor may also underlie the data at each grade level. For example, the amount of total variance explained by the first initial factor extracted at grade 8, 1996 was 42%. The amount of common variance explained by the first initial factor extracted at grade 8 was 82%. In addition, the inter-factor correlations among the four rotated factors (Table 14-2) were moderately high, again reflecting the general achievement factor.

Results of the above factor analyses provide favorable evidence with regard to construct validity. Considering that open-response items yield scores included as components of an overall accountability index used in a high-stakes accountability system, it is good that the items from different content areas hold together as reflected in the inter-factor correlations and the amount of total/common variance explained by the first initial factor extracted. In addition, it is positive that content factors were found across KIRIS grades and years. The existence of such factors suggests that each content area as measured by open-response items is measuring a construct discernible from the others. Of course, the present analyses provide an initial exploration of the factor structure of KIRIS open-response items. Because KIRIS is a complex, multi-faceted assessment system, it could be informative to use confirmatory methods to develop and test more comprehensive models exploring the construct validity of KIRIS.

TABLE 14-1
GRADE 8 STUDENT-LEVEL FACTOR ANALYSIS (N=50,321) USING OPEN-RESPONSE COMMON ITEMS - ROTATED SOLUTION

Test		Factor Pattern Matrix (Standardized Regression Coefficients)			
Subject	Item Number	I	II	III	IV
Reading	1	**.65**	.01	.04	-.02
	2	**.65**	.01	.02	-.05
	3	**.60**	.00	.01	.03
	4	**.59**	.05	.05	.07
	5	**.57**	.04	.07	.08
	6	**.50**	.05	.05	.09
Mathematics	1	.16	**.45**	.05	.04
	2	.02	**.62**	.02	.04
	3	.03	**.61**	.03	.05
	4	.02	**.58**	.04	.06
	5	-.03	**.59**	.03	.00
	6	.03	**.62**	.02	.03
Science	1	.06	.12	.06	**.47**
	2	.03	.18	.05	**.40**
	3	.02	.05	.05	**.50**
	4	.10	.07	.07	**.46**
	5	.08	.02	.08	**.46**
	6	-.02	.07	.05	**.50**
Social Studies	1	.05	.10	**.52**	.08
	2	.08	.02	**.61**	.02
	3	.09	.05	**.59**	.00
	4	.03	.05	**.61**	.05
	5	-.04	.01	**.65**	.03
	6	.07	-.01	**.62**	.03

TABLE 14-2
GRADE 8 STUDENT-LEVEL FACTOR ANALYSIS INTER-FACTOR CORRELATIONS

	I	II	III	IV
I	1.0			
II	.61	1.0		
III	.66	.62	1.0	
IV	.65	.69	.67	1.0

CONCURRENT VALIDITY-RELATED EVIDENCE

Another measure of validity is how well a test correlates with accepted measures of the same or similar constructs. To the extent that few or no other "primarily performance-based" assessments exist for comparison with KIRIS, options are limited for demonstrating concurrent validity--even though many traditional measures have been enhanced to include constructed-response supplements in conjunction with multiple-choice or selected-response items. The best one can do is to compare the performance of students on KIRIS to accepted or "traditional" tests of academic achievement, despite the fact that they will not assess exactly the same construct as KIRIS. In this case, one hopes that the correlations will be high, since much of what is measured by most testing programs will be in common, but not too high. If the correlations are very high, it would imply that whatever is supposed to be new in KIRIS is either missing or highly correlated with traditional forms of assessment. If that were the case, use of KIRIS could not be justified on the basis of what it measures, since the traditional forms of assessment are usually so much less expensive and time-consuming to administer.

Concurrent validity can be assessed through correlational study at the student level, the level at which most achievement measures traditionally have been pitched, or at the school level, the level at which KIRIS is pitched. State-level comparisons can also be made.

STUDENT-LEVEL RELATIONSHIPS. In addition to providing concurrent validity evidence, a good reason for comparing KIRIS to traditional forms of assessment is that those traditional measures are still in use. Tens of thousands of Kentucky high school students will take the ACT each year for some time to come for college admissions. If KIRIS proved to be uncorrelated with that measure, it would place students, parents, and teachers in the uncomfortable position of having to choose the test on which they would like to focus their attention. Fortunately, as shown in the KIRIS Accountability Cycle I Technical Manual, there is significant correlation between KIRIS and ACT scores, and it is not overly presumptuous to assume that increased learning that leads to improvement on one is likely to lead to improvement on the other. For 1993-94, KIRIS raw score and ACT scaled composites correlated .74, with mathematics, reading and science scores correlating .67, .47 and .46, respectively.

Comparing KIRIS with measures used widely in the past can give a sense of continuity and relationship, in addition to providing concurrent validity evidence. The Comprehensive Test of Basic Skills (CTBS$_4$) was given widely in Kentucky before the inception of KIRIS, and it would be disquieting if there were little or no relationship between the two measures. Again, as shown in the KIRIS Accountability Cycle I Technical Manual, there is a high degree of relationship, with CTBS Normal Curve Equivalents and 1991-92 KIRIS raw scores correlating .61 and .63 for mathematics at grades 4 and 8, respectively, and .47 for reading at both grades. For those who still question how *avant garde* KIRIS is, it should be reassuring to them to know that the correlations between the old and new are sufficiently high to show that while the KIRIS tests are a step in a new direction, they clearly are not too far a leap from traditional testing.

SCHOOL-LEVEL RELATIONSHIPS. Considering that KIRIS scores are used for school accountability, it may be argued that a high degree of relationship between KIRIS and other test scores obtained by schools is even more essential evidence of concurrent validity than correlations among student- level scores. Obtaining evidence of this kind is problematic insofar as few other tests are administered to all students in a school, in contrast to KIRIS, which was given to about 99% of all students in most years of the second accountability cycle. (For KIRIS, the remainder are excluded typically for medical or language reasons, or participate in school accountability through the alternate portfolio program, as noted in Chapter 7.) And of the other tests not given to all students, very few are administered to a representative or even approximately random sample of students at participating schools, further diminishing the meaningfulness of school-level scores. The best available comparison in this regard is the National Assessment of Educational Progress (NAEP), as administered in reading in the 1992 trial state assessment program. As shown in the KIRIS Accountability Cycle I Technical Manual, there is a high degree of relationship between school-level mean 1992 NAEP scaled scores and 1991-92 KIRIS raw scores; they correlate .74, .78 and .79 at grades 4, 8, and 12, respectively. This level of relationship was obtained despite the fact that NAEP score school means for this study were based on a limited number of students (9-31 students for grade 4, 16-31 for grade 8, and 14-45 in grade 12, with the exception of one school that tested 87 students). Note also that school-level but not student-level matching information was available for NAEP scores, so that school-level scores represent the most fine-grained level of comparison possible between scores on KIRIS and NAEP.

STATE-LEVEL RELATIONSHIPS. Considering that KIRIS is administered only in Kentucky, there are no other states with which to compare student performance on KIRIS. However, it is possible to compare changes in state-level scores on KIRIS over time with state-level changes over time on other measures. The challenge in making such a comparison is that, relative to the first year of testing, some improvement in KIRIS scores is likely to occur as a result of directing school curricula toward the test and familiarizing students with responding to open-response questions in general. Initial gains from the 1991-92 baseline are unlikely to generalize to other tests, but are a predictable, initial result of implementing a high-stakes testing program. To the extent that this effect has been observed with multiple choice tests[7] used in a high-stakes setting, a finding that initial KIRIS gains do not generalize to other tests is not evidence against score validity, but rather an indication that caution must used in interpreting score gains relative to the first year of high-stakes testing.

At grades 4 and 8, NAEP scores could provide a comparison for reading, mathematics and science score gains, except that NAEP is limited in this regard to comparisons of mathematics scores, involving the years 1996 and 1992. Science results for NAEP were not released for publication in time for inclusion in this chapter, and there was no state-level NAEP reading testing in 1996. However, in mathematics at both grades 4 and 8, Kentucky students achieved gains on NAEP. Gains were also observed in KIRIS mathematics scores over the course of the first accountability cycle, 1991-92 to 1993-94, with some gain over the course of the second accountability cycle as well. This is favorable, but limited evidence. At the same time, NAEP probably represents the most appropriate comparison with KIRIS.

Regarding high school student performance, there is not a measure such as NAEP for comparison with grade 11 KIRIS scores. The most widely taken test at grade 11 other than KIRIS is the ACT, for which scores were flat over the course of the second accountability cycle. So, too, were KIRIS scores in the last three years of the second accountability cycle. However, the last three years yielded KIRIS scores higher than 1992-93, the second year of accountability, and one in which particularly low scores were obtained by grade 12 students. These scores were attributed, albeit somewhat anecdotally, to a lack of student motivation, and led to the movement of high school KIRIS testing to grade 11. This is mixed rather than clearly favorable or unfavorable evidence with regard to validity. The quality of this evidence is not as strong as would be desired, based on differences in the construction and format of ACT relative to KIRIS, and on the fact that ACT takers represent a self-selecting sample of Kentucky youth. In order for the comparison to be at all reasonable, the self-selection process must be similar in each year, as must the stakes experienced by and other motivation of KIRIS and ACT-takers. Because cross-year comparisons of ACT and KIRIS scores logically cannot reflect changes in the KIRIS performance of students who do not take the ACT, it is better to limit score comparisons to those of students who take the ACT and KIRIS over the course of the second accountability cycle. Matching data of this kind were not available for the latter half of the second accountability cycle at the time this report was released, but will be assembled for future research.

---

[7] See, for example, Linn, R. L. (1995). Assessment-based reform: Challenges to educational measurement. Educational Testing Service: Princeton, NJ.

Note that there are methodological issues related to scaling in making comparisons across measures. Not only is each test built to its own specifications, but each measure has its own scale. As long as each measure provides an indication of whether changes over time are statistically significant, it is possibly to compare trends broadly. Comparing the magnitude of changes on one measure with magnitude of changes on another is more complicated, especially when multiple sets of scores are available for one or the other of the measures (such as theta and standards-based--Novice, Apprentice, Proficient, Distinguished-- scores on KIRIS open-response tests).

CRITERION-RELATED VALIDITY EVIDENCE

A vast potential source of validity evidence to support or refute the inference that accountability gain scores reflect improvements in school performance is schools themselves. The primary challenge associated with taking advantage of this rich source of information are that it is logistically difficult (and therefore expensive) to gather meaningful data on schools. A lesser challenge (and, some would argue, a potential advantage) is that information of this nature does not necessarily lend itself readily to quantification, so that results must be considered mostly in qualitative terms.

Through its subcontractor, KDE has undertaken qualitative study of schools with varying levels of improvement on KIRIS, but data are not yet available for presentation. However, a case study of 16 schools conducted by Carolyn Kelley[8], a senior research associate at the Wisconsin Center for Educational Research, provides important initial evidence in this regard. She found that successful schools had taken specific actions to achieve success, including analyzing test results to identify weaknesses, setting goals, changing curriculum and using professional development effectively. By contrast, low-success schools had not changed their curriculum and had not used professional development to learn about the new learning goals. These are favorable results for KIRIS, but more data is necessary. Additional evidence will be gathered to study the relationship between KIRIS score gains and changes in school practice.

---

[8] Kelly, C. and Protisik, J. (1997). Risk and reward: Perspectives on the implementation of Kentucky's school-based performance award program. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

CHAPTER 15
CONSEQUENTIAL VALIDITY OF KIRIS AND THE KENTUCKY ASSESSMENT PROGRAM

CONSEQUENTIAL VALIDITY

An important aspect of an assessment's validity is whether it is used according to its intended purposes, or more broadly, whether its effects, or consequences, are aligned with its goals. This attention to "consequential validity" has been articulated most forcefully by Messick[1]. Messick links the *interpretation* of what a test represents or means (construct validity) with the *use* of a test or test information (consequential validity) as two essential aspects of validity. Consequential validity thus raises the question of the extent to which the Kentucky assessment program is achieving its goals and not having unintended, negative side effects.

INTENDED GOALS OF KENTUCKY ASSESSMENT PROGRAM

The role of the Kentucky assessment program, including the KIRIS assessments, is to promote educational improvement for all children in the state. It does this in three major ways:

1. The assessment program provides goals, standards, and criteria for educational achievement. These goals, standards, and criteria are found linked together throughout the assessment program. They include the statements of goals in the KERA legislation, the Academic Expectations, and the links between Academic Expectations and specific items and their scoring guidelines. The assessment program also includes operational definitions of success, various performance levels, relative weights of assessment components, and so on.

2. The assessment program provides useful information on progress towards achieving those goals made by schools. Although the major informational use of KIRIS scores are in relation to school accountability as mandated in the KERA legislation, much assessment information is also provided that is useful for monitoring achievement and progress of individual students, the state as a whole, and various subgroups within the state.

3. The assessment program provides useful information on potential differential impact of the assessment program in particular, and of education within the state, by region, and on various subgroups such as ethnic and racial minorities, children living in poverty, and so on.

---

[1] Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement (3rd Ed.)*. New York: ACE/MacMillan.

EVIDENCE AND INTERPRETATION OF CONSEQUENTIAL VALIDITY

The KERA legislation and subsequent programs, including the assessment program, have engendered much discussion and activity. However, relatively little formal research is available on the impact of the assessment program on classroom practice, teacher development, or support of educational reform in Kentucky. In addition, in the available research it is often difficult to separate effects of the assessment program from other aspects of educational reform. For example, while the criteria for mathematics portfolios have made an impact on classroom instruction as a result of being included in the assessment system, some changes have resulted from the fact that Kentucky has adopted standards that effectively reflect the national standards promulgated by the National Council of Teachers of Mathematics.

It should be noted that the discussion in this chapter of the consequential validity of the Kentucky assessment program paints broad strokes, which may not apply to every classroom in every school. Change is taking place at different speeds and in different forms throughout the Commonwealth, often for different combinations of reasons. In addition, the assessment program, and people's understanding of it, has changed over time. Continuing research will be required to provide more complete results of the consequential validity of the Kentucky assessment program, as well as to keep research findings up to date.

CONSEQUENCES: PROVIDES GOALS, STANDARDS, CRITERIA FOR INSTRUCTION AND CURRICULUM

Evidence continues to accrue regarding the effects of KIRIS on instructional practice, teachers' professional development, and support for educational reform. Evidence presented in the KIRIS Accountability Cycle I Technical Manual addressed impact on instructional practice, professional development, and educational reform, citing several studies.[2] Taken as a whole, these studies suggest that KIRIS has already had an impact on instructional practice, particularly with regard to portfolios and/or results of participation in professional development activities, and that the assessment met criteria necessary for contributing to educational reform in science and mathematics.

Since the completion of the KIRIS Accountability Cycle I Technical Manual, several other studies have taken place that address consequential validity in the context of curriculum and instruction. In lieu of attempting to summarize all research in this regard, it is worthwhile to note the results of one of the most prominent ones. A survey- and interview-based RAND study[3] found that "about 40 percent of the teachers reported that the open-response items and portfolios have had a great deal of positive effect" on instruction, with only half as many teachers endorsing this view of performance events, and almost none considering multiple-choice items to have had such a positive effect. Note that the study examined a broad range of perceived effects, including perceived negative impact on instruction, and provides a more extensive discussion of its findings than can be afforded in this chapter.

---

[2] Appalachian Educational Laboratory. (Dec., 1994). Instruction and assessment in accountable and non-accountable grades. *Notes From The Field*, 4(1), 1-2.; Pankratz, R., Ochs, D. et al. (April, 1995). *Configuration maps: Results from Kentucky*. Papers presented at the annual meeting of the American Educational Research Association, San Francisco, CA; Policy Studies Associates, Inc. (1994). *Third-year evaluation of the nine-site program initiative*. (A report to the U.S. Department of Education.) Washington, DC: Author; McCollum, H. et al. (August, 1994). *Portfolio assessment in mathematics: Views from the classroom. Annual report*. Washington, DC: Policy Studies Associates, Inc.;Roberts, R. & Key, S. (September, 1993). *Kentuckians' expectations of children's learning: The significance for reform*. A public report prepared for the Prichard Committee for Academic Excellence and the Partnership of Kentucky School Reform, Lexington, KY: Roberts & Key, Inc. (Available from the Prichard Committee for Academic Excellence, P.O. Box 1658, Lexington, KY 40592-9980.); Winograd, P., Jones, D., & Perkins, F. (submitted). *The politics of alternative assessment: Lessons from Kentucky*. (Manuscript obtained from first author. ; Laguarda, K. G., Breckenridge, J. S., Hightower, A.M., & Adelman, N. E. (September, 1994). *Assessment programs in the statewide systemic initiatives (SSI) International, primary contractor.)* Prepared under contract for the National Science Foundation, SRI International, primary contractor. Washington, DC: Policy Studies Associates, Inc.

CONSEQUENCES: PROVIDES INFORMATION ON STATUS AND PROGRESS

KIRIS provides information to schools and districts in several forms. These are described more fully in the chapter on Reporting in this Technical Manual. Essentially, KIRIS reports include scores on each subject matter area and a non-cognitive index for schools and districts each year. KDE/ASME produce a biennium report that summarizes each school's performance in terms of the baseline and two subsequent years of the first accountability cycle, and summarizes the school's status in relation to rewards and assistance. Student reports are produced each year for subject matter areas, including portfolios. The student reports are sent by KDE/ASME to schools, where schools decide how to distribute them. In addition, each year KDE/ASME produces a summary report for the state, by region, by gender and by ethnic group.

Schools, districts, and classroom teachers have reported using the score reports in a variety of ways consistent with the intent of KIRIS. The most common use is in broad program review to mark progress over a year or two, and focus resources for instructional program improvement. Analysis of KIRIS scores comprises an essential part of every school's annual Transformation Plan. However, KIRIS scores have been used by schools and teachers for other purposes. There are increasing requests to have KIRIS provide information in addition to the school accountability function it was originally designed to provide. There have been wide-spread calls for additional information in the other traditional evaluation areas:[4]

1. for example, individual student achievement status, for use on report cards;

2. individual student comparative status, or for college admissions;

3. school comparisons (the news media routinely convert KDE/ASME reports into "rankings" that facilitate comparisons between schools and districts);

4. student diagnostic information for monitoring student progress and informing instructional changes by classroom teachers;

5. instructional program evaluation, to monitor and improve instructional programs, school curricula, and inform teacher assignments and professional development.

---

[3] Koretz, D. M., Barron, S.,Mitchell, K. J., & Stecher, B. M. *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)*. Sante Monica, CA: RAND.

[4] For example, see the report done by The Evaluation Center, Western Michigan University, (January, 1995). *An independent evaluation of the Kentucky Instructional Results Information System (KIRIS)*.

While KIRIS currently provides results that address these areas to some extent, most would require substantial changes in KIRIS design and or operation. Some of these uses are currently possible; some are possible but somewhat incompatible with the intended uses of KIRIS results. For example, using student KIRIS results as the sole basis for school report card grades is viewed by KDE as an inappropriate use of KIRIS results. Providing diagnostic information for individual students would require much more rapid feedback than is currently possible, as well as local interpretation. KDE holds that such diagnosis is more appropriately undertaken by classroom teachers using classroom assessments other than, or at least in addition to, KIRIS assessments, and is more appropriately undertaken earlier in the school year (i.e., earlier than April).

CONSEQUENCES: IS FAIR TO SCHOOLS

It is important that the Kentucky accountability program provide a fair challenge to all schools. This is especially true regarding the consequences of rewards and assistance based on the assessment results. As noted in the KIRIS Accountability Cycle I Technical Manual, several factors have been examined to explore whether the Kentucky accountability program is fair to schools. The factors include geographical location of school, racial/ethnic composition, economic status of students, initial baseline score, school size, and grade level organization. Based on these analyses for the first accountability cycle, the Kentucky assessment program appeared to be fair in that rewards and assistance were distributed across these dimensions without statistically significant unevenness. The exception was grade level, where proportionally more elementary schools received rewards than did middle schools or high schools. This result seemed to be explained by differences in education at these grade levels, rather than bias. Results for the second accountability cycle were similar.

PROGRAM-SPECIFIC SCHOOL-LEVEL EFFECTS. Beyond being fair with regard to characteristics of student enrollment, KIRIS should not disadvantage schools participating in programmatic efforts to improve student learning. The effort in which Kentucky schools participate most widely is Title I, a federal program established to serve economically disadvantaged students by providing supplemental funding to schools, based on the poverty level of students in the district and school. Between the 1992-93 and 1995-96 school years, about 70% of the public schools in the Commonwealth participated. Kentucky is unusual in the nation: as a high poverty state, Kentucky has over half its schools with a reported 50% or more students qualified for free or reduced price school lunch.

Over the course of the second accountability cycle, elementary, middle, and high schools that participated in the Title I program achieved relatively greater progress toward their KIRIS improvement goals than did non-Title I schools. This is a favorable finding with regard to consequential validity, or at least serves to balance the opposite occurrence in the first accountability cycle, as described below.

In Cycle 1, the elementary schools, on average, achieved 115% of their KIRIS goal. Non-Title I elementary schools achieved 126% of their KIRIS goal while Title I schools achieved 113% of their goal. Although the percents of goal achieved by all schools are all above 100%, non-Title I schools in Cycle 1 exceeded their improvement goal by substantially more than did Title I schools. However, in Cycle 2, the reverse occurred: on average, elementary Title I schools made relatively greater growth (129% of their KIRIS goal), than did non-Title I elementary schools (119% of their goal). This overall growth and the reversal in the relative growth pattern in Title I and Non-Title I schools shows that elementary schools participating in Title I programs exceeded their KIRIS goals to a greater degree than did non-Title I schools over the past four years (1992-96). Also, on average, not only are elementary schools achieving over 100% of their KIRIS goals, but they have substantially increased their rate of growth from Cycle 1 to Cycle 2. These results indicate that the Title I program in conjunction with other educational reform efforts, is having a positive effect on elementary education in Kentucky.

The pattern of relative achievement of middle schools is similar to that of elementary schools. The Title I schools were relatively higher than non-Title I schools in Cycle 2 in terms of achieving their improvement goals. In Cycle 1, non-Title I schools achieved a greater percentage of their goal than did Title I schools, although not by a substantial margin. Non-Title I middle schools achieved 85% of their KIRIS goal whereas Title I schools achieved 82% of their goal. In Cycle 2, the reverse pattern occurred: on average, Title I middle schools achieved 57% of their KIRIS goal, whereas non-Title I schools obtained only 46% of their goal.

Only 43 of approximately 200 high schools (and a total of only 776 Grade 11 students) participated in Title I in the 1995-96 school year. Because so few high schools and their students participated in Title I programs, analyzing changes in school performance based on Title I status should be done with extreme caution. The pattern of relative achievement in high schools is similar to that of the elementary and middle schools. Title I schools achieved a greater percentage of their improvement goal than did non-Title I high schools in Cycle 2.

CONSEQUENCES: IS FAIR TO STUDENTS

Given that KIRIS is designed to foster learning for all the children in the Commonwealth, it is important that the assessment and the educational system be fair to all students. One crucial fairness question in any testing program is whether the assessment adversely affects one group of students compared with another. If differential performance exists based on race and/or gender, additional investigation is warranted into the assessment or the educational system.

DEMOGRAPHIC COMPOSITION OF KENTUCKY SCHOOLS. Typically, student performance is compared between males and females, and among ethnic groups. Demographically, Kentucky's school population has been extremely stable over the accountability cycle. The distribution of male and female students is relatively even; however, the racial/ethnic mix of students in the Commonwealth is not. Mirroring the state's racial mix, African-Americans (constituting about 9% of the population) and whites (comprising about 89%) are the two major racial groups in these grades; all other ethnic/racial groups comprise only 2% of students tested in grades 4, 8, and 11. Given the small numbers of Hispanic, Asian, and Other students, differential racial performance analysis is limited to differences between white and African American students.

MEASURES OF STUDENT PERFORMANCE. As noted elsewhere in this Technical Manual, the KIRIS assessment yields two student-level scores. One is theta, a measure of ability within Item Response Theory (IRT) created by modeling the patterns of student responses to test questions using a graded response model (Samejima 1969). A continuous variable usually standardized so that its range generally runs from about -3 to +3, theta is highly correlated with the total raw score on constructed response (also call open response) subtests. Another measure of student performance for each subtest that can be created from the KIRIS assessment is the constructed response index (or performance level index) which is derived using cutpoints on the theta scale selected to represent performance standards, and assigning weights of zero, 40, 100, and 140 to the performance levels Novice, Apprentice, Proficient, and Distinguished, respectively. These subtest indices are an alternate measure of student performance. Like the theta scores from which they arise, the constructed response subtest indices can be averaged to create a composite index. This chapter will focus on theta, although a more detailed report which includes results using both theta and the constructed response index is available from the Kentucky Department of Education, Office of Curriculum, Assessment and Accountability.

PERFORMANCE DIFFERENCES: DO THEY EXIST? Figures 15-1 through 15-3 present subgroup performance by race and gender for the entire accountability cycle.[5] These figures reveal several things about performance on the KIRIS assessment. First, the data show that student performance for all groups has improved over the course of Accountability Cycle 2. Second, performance differences are evident within the data. White females have the highest average performance followed, generally, by white males and African-American females. The lowest performance in each grade is turned in by African-American males. Although the performance differences between groups vary in size according to the content area and the measure used, these patterns generally hold for both the subtests and the composite measures in all three accountability grades.[6]

PERFORMANCE DIFFERENCES: HOW BIG ARE THEY? Figures 15-1 through 15-3 provide a pictorial view of performance difference over time, but offer limited understanding of how large the performance differences between males and females and between African Americans and whites are. There are two ways of estimating the size of the difference between subgroups: percentiles and effect size statistics. In the first method, percentiles divide white student scores into 100 equal parts and assess the percentile at which the median black student score would fall. The other estimate, the effect size statistic, expresses performance differences in terms of standard deviations much like z-scores. The z-score statistic relates the position of a particular student score to the number of standard deviations it is from the average student score. Here, instead of comparing a particular student score, a subgroup mean is substituted for comparison with another subgroup's mean in terms of its standard deviation. In particular, the size of performance differences for gender was calculated by dividing the difference between the female and male means by the standard deviation for female. Similarly, the performance difference for race was calculated by dividing the mean difference between whites and African Americans by the standard deviation for whites.

Results of analysis using percentile ranks and effect size statistics show that the gap between males and females is smaller in the fourth grade than the performance gap at the eighth and eleventh grades, especially in reading and social studies. It is possible that the smaller gap in fourth grade scores is due in part to KERA's changes to the educational structure since the children tested in the fourth grade in 1996 are the first children to have full benefit from KERA's primary grade restructuring. Only future research will be able to disentangle the age and educational structure effects.

---

[5] Performance differences for content area sub-tests are available from the Kentucky Department of Education, Office of Curriculum, Assessment, and Accountability.

[6] There are two noteworthy exceptions to this recurrent pattern. In Grade 4 mathematics, African-American male performance surpassed African-American female performance for the last two years. Also, in Grade 8 reading, African-American female performance pulled even with white male performance over Accountability Cycle 2.

FIGURE 15-1

KIRIS Composite Theta Scale for Grade 4 from 1993 to 1996 by Gender and Race

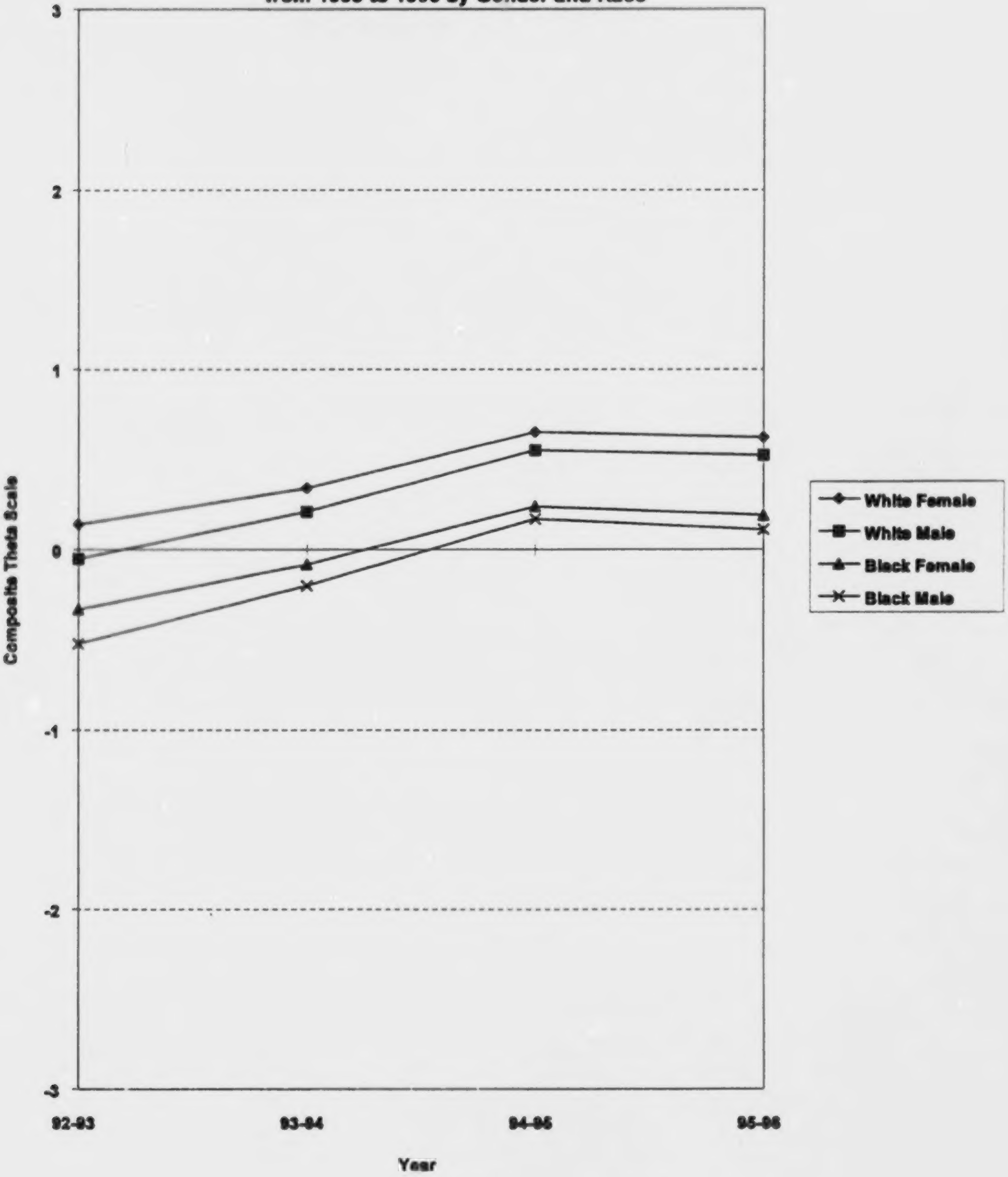
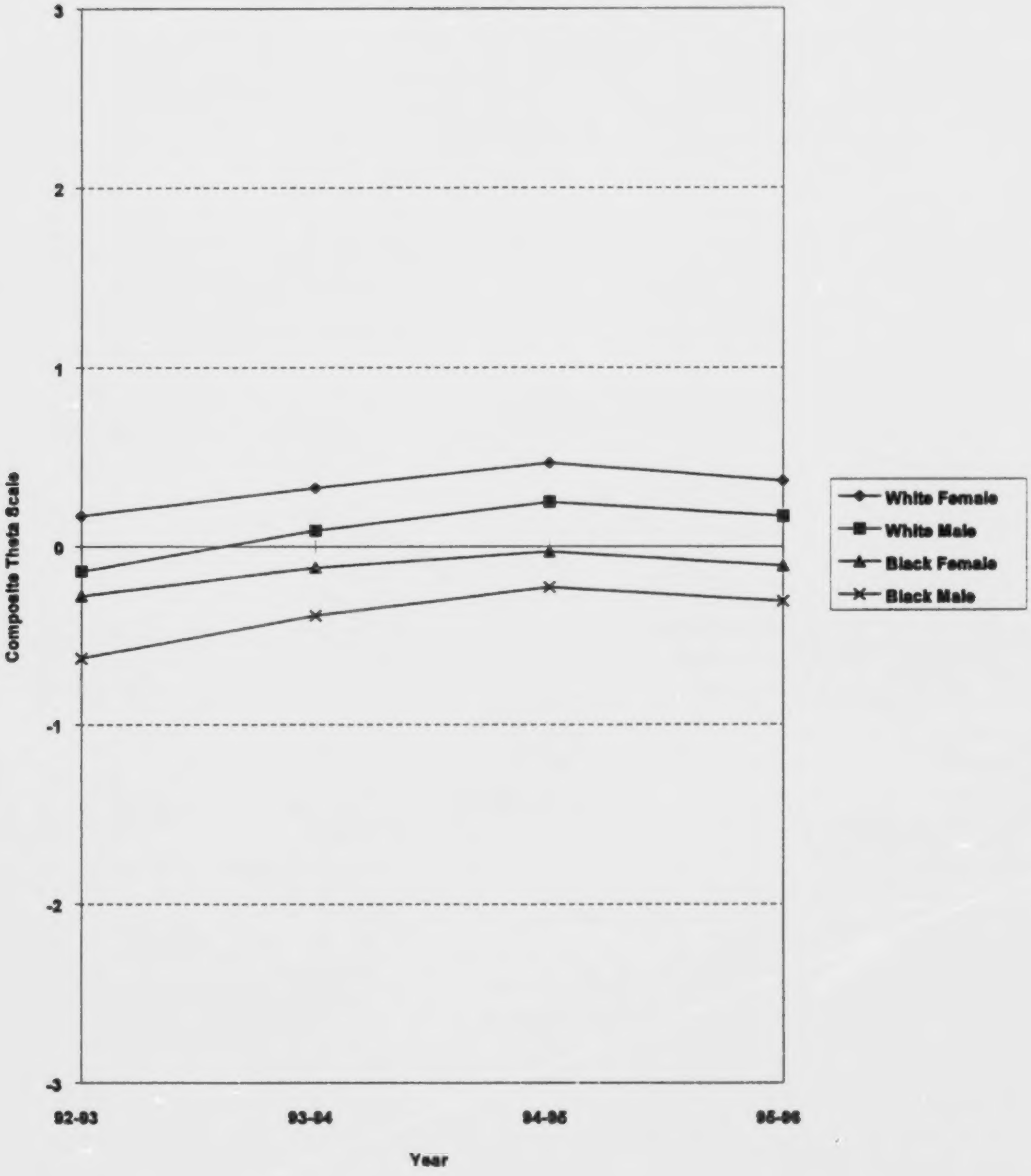


FIGURE 15-2

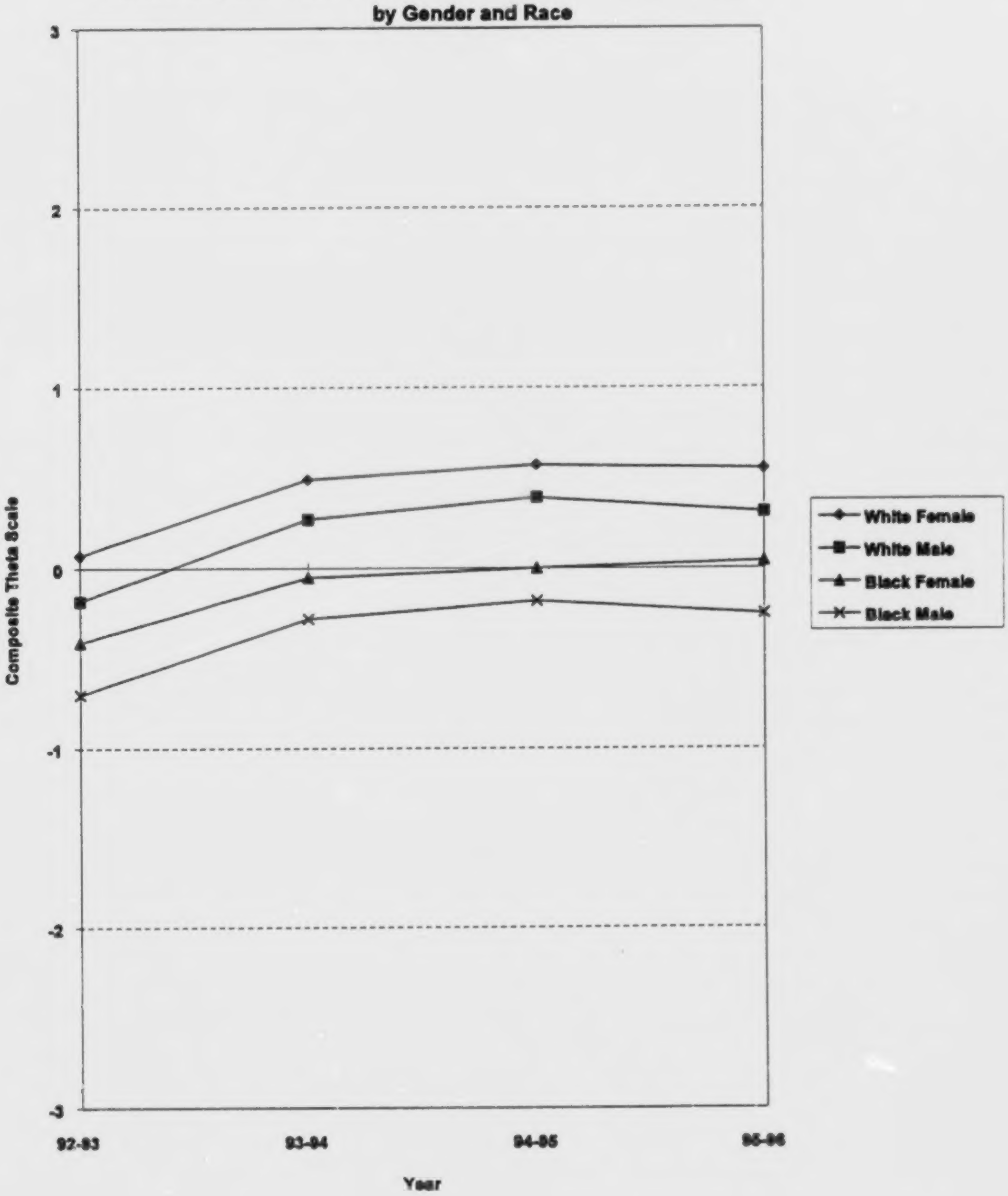
KIRIS Composite Theta Scale for Grade 8 from 1993 to 1996 by Gender and Race



204

FIGURE 15-3

KIRIS Composite Theta Scale for Grades 11/12 from 1993 to 1996 by Gender and Race



Examination of the performance gap between African-American and white students clearly shows that the gap between African-Americans and whites is much larger than that between males and females. According to the percentile rank analysis, African-American scores range 20 to 28 percentile points below whites on any given subtest in 1995-96. Therefore, the median African-American child scores about one-half to three-quarters of a standard deviation below the median white child. Like the gender gap, the racial performance difference is smallest at Grade 4; however, unlike the gender gap, the difference between white and black students is smallest in reading and social studies.

PERFORMANCE DIFFERENCES: INCREASING OR DECREASING? Using the statistics available to judge the size of performance differences, it is possible to assess whether these performance gaps have increased or decreased over the accountability cycle. Analyses of change over time show that the gap between males and females has decreased in most areas for most grades. The exceptions to this pattern are differences in reading at Grade 8 and reading and science at Grade 11. However, even in areas where the gender gap increased, the increase was very slight (less than a tenth of a standard deviation).

Analyses of changes in the size of the performance difference between African-American and white students show that the performance differences between white and African-American students widened over the accountability cycle. While the increase in the gap was very slight at grade 4, the change in performances of African-Americans and whites in grades 8 and 11 was more pronounced. In mathematics, science, and social studies at grade 8 and mathematics in grade 11, the difference increased by a fifth to a quarter of a standard deviation in four years.

POSSIBLE EXPLANATIONS FOR RACIAL/ETHNIC PERFORMANCE DIFFERENCES. Although there is a sizable difference in performance on KIRIS between whites and African-Americans, this does not necessarily mean that the KIRIS assessment or the accountability system is biased. The KIRIS assessment is reviewed extensively each year for possible bias during test construction. Items that are judged by the bias committee to disadvantage student subgroups are either dropped from the test or modified to remove the potential source of bias. Additionally, examination of the available data from NAEP and ACT reveals a similar performance gap between white and African-American students (Education Week reports that the race gap decreased from 1970 to 1988 and has been increasing ever since [Olson, 1997]). These differences are roughly comparable to those reported by KIRIS, making support for the idea that the KIRIS assessment is biased more difficult to sustain.

Given the existing safeguards and the comparability of results from other testing programs, bias in the KIRIS tests alone is probably the last explanation that should be considered for differential performance. Instead, research should focus on individual, community, school, or district level phenomena to that might contribute to the observed differences. Unfortunately, much of the data necessary for investigating these possible effects thoroughly (particularly at the individual level) is not always available to the Kentucky Department of Education. However, the following analyses of the effects of the geographic distribution of minorities within the state and the relationship between a school's accountability index score, its racial mix and its socioeconomic status, represent ongoing research into observed performance differences.

GEOGRAPHIC DISTRIBUTION OF RACE. African-Americans in Kentucky are not evenly spread throughout the Commonwealth's 120 counties, but instead are located primarily in the major cities and near military bases within the state. Such an uneven concentration of minority students may profoundly influence the size of the gap between

African-Americans and whites. To examine this possibility, KDE researchers have examined the effects of district and school racial mix on the racial performance gap.

To examine the possibility that students in school districts with higher percentages of minority students enrolled differ in their achievement from students in school districts with lower minority percentages, the students tested in grades 4, 8, and 11 were divided in two groups based on whether their district was above or below the state average percentage minority enrollment (11%).[7] Similarly students were also divided based on whether their school had above or below the state average minority enrollment. Percentile and effect size statistics of the performance gaps were then calculated for students in each division.

The results of this analysis show that the performance gap between African Americans and whites varies with the racial mix of the district and school. First, the size of the performance gap between white and African-American students is larger for students in the thirty-three districts with above the state average of percentage minority enrollment than it is for students in districts with below the state average. The performance gap is also larger between African-American and white students in schools with above average percentages of minority students. Second, the gap increases in size from Grade 4 to Grade 11. Third, an examination of the interaction between the racial mix of the school and the district shows that students in districts with larger than average percentage of minorities attending schools with above the average percentage of minority students display an even larger gap than would be expected based on the differences between schools and between districts alone.

One interpretation of these results critiques school districts with high percentage minority enrollment, suggesting that African-American students in these districts are not served well compared to those who go to school in less diverse districts. Another suggests that the gap between white and African-American students in rural school districts is smaller because whites in less diverse school districts do not perform as well as whites in districts with higher percentages of minority students. Subsequent analysis of white performance shows that whites in the above average percentage minority enrolled school districts score lower than whites in the other school districts in all grades tested, removing support for the second explanation.

RACE AND SOCIO-ECONOMIC STATUS. Another frequently advanced hypothesis suggests that African-American student performance lags behind white performance as a function of poverty. Because African-American students are more likely to reside in low income neighborhoods within the Commonwealth, the effects of their community's economic status (i.e., the percentage of households in poverty) becomes confounded with the effects of the racial mix of the school. A school level analysis reveals several interesting things about a school's achievement, its racial mix, and its socioeconomic status. First, it is readily apparent that a school's racial mix and its socioeconomic status have more impact on a school's achievement in a single year than over an accountability cycle. Still, even though a school's poverty level and its minority enrollment may be confounded, there is a statistically significant school level racial mix effect above and beyond the effects of school poverty in Grades 4, 8, and 11 when examining achievement in a single year and a very small, although statistically significant, effect of a school's racial mix on the percent of school's goal achieved over Accountability Cycle 2. While these school level findings cannot be used to infer information about individual student performance, the results take support away from the idea that the effects of a school's racial mix can be explained away as an effect of its socio-economic status. Note that the Kentucky Department of Education does not have access to individual student economic data, so that examination of the effects of poverty is presently limited to school-level study.

---

[7] The school and district level analysis uses percent of minority enrolled in the accountability grades instead of percent of African-Americans enrolled; however, these two measures are comparable given the small number of other minority students.

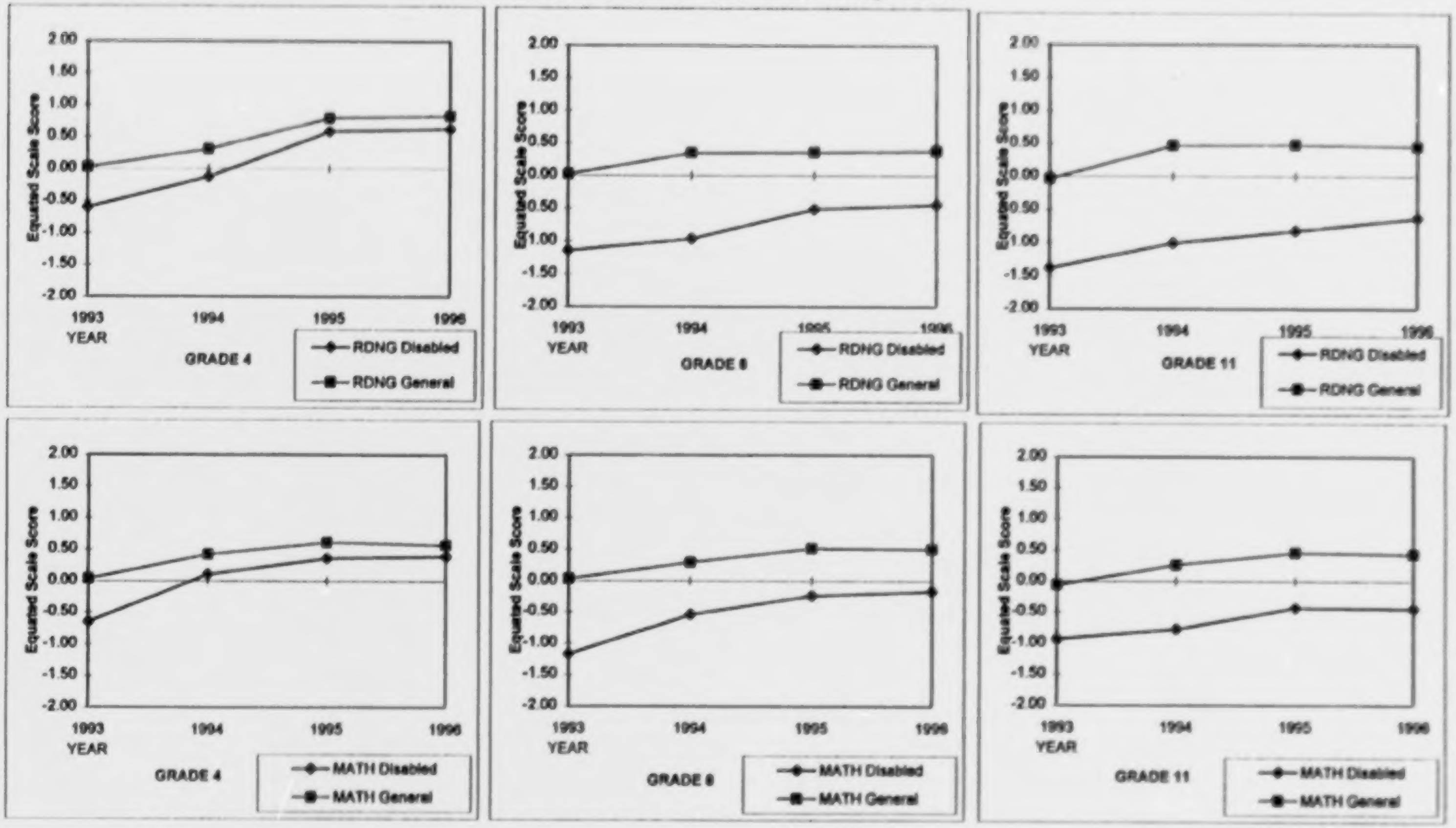
STUDENTS WITH DISABILITIES. Kentucky's policy is that schools are accountable for *all* students, with very minor exceptions. This accountability policy is reflected in the inclusion of students in the KIRIS student assessment and school accountability program: more than 98% of students in the accountability grades are included in the assessment. Because less than 1% of these students participate in school accountability by taking the Alternate Portfolio, one implication of this policy is that the assessment includes many students with disabilities. Kentucky is one of two states in the nation that has such an inclusive state assessment policy regarding students with disabilities, according to the National Center on Education Outcomes for Students with Disabilities. Consequences of this use of KIRIS can be considered by analyzing the KIRIS scores of students with disabilities, many of whom receive accommodations on the assessment.

Under federal law, students with disabilities may go through a process by which appropriate accommodations are identified and provided in their instructional settings. These accommodations are intended to help the student learn in the least restrictive environment, and to allow the student maximally to develop and demonstrate his/her capacities. An example is a blind student who, with the use of Braille print, a computer, a human reader, or other assistive technologies or accommodations, can learn and perform well. The Kentucky state assessment policy seeks to mirror the non-restrictive instructional environment mandated by law, in that the student may have access to the same accommodations on the KIRIS assessment as are regularly provided for in the classroom, as specified in the student's Individual Educational Plan (IEP) or 504 plan, which are the federal mechanisms for documenting learning goals, needs, and accommodations. These adaptations/accommodations are categorized as follows:

None -- No adaptations provided. The student with disabilities takes the assessment under the same conditions as other students.
Oral -- The KIRIS assessment is read to the student.
Dictation --The responses to KIRIS assessment are dictated by the student and written down by a scribe.
Paraphrasing -- The content of the KIRIS assessment is paraphrased for the student.
Technological -- Technology normally used by the student is made available during the KIRIS assessment, e.g., large print or Braille versions of the test for visually impaired students.
Other -- Other adaptations/accommodations made available in the delivery of instruction are made available during the KIRIS assessment.

Achievement on the KIRIS assessment for the general population (GP) and students with disabilities (SWD) can be compared in terms of mean scale scores of the reading, mathematics, science and social studies KIRIS On-Demand components of the assessment administered during the Accountability Cycle 2--school years 1992-93 through 1995-96. For Grades 4, 8, and 11, the average achievement (theta) scores for students with disabilities were lower than the average scores for the general population. Figure 15-4 presents results for mathematics and reading at these three grade levels; similar results were obtained in science and social studies. Overall, accommodations appear to be more effective and extensive for SWD at the Grade 4 level than at grades 8 and 11. There were different patterns for each grade in terms of the amount of difference and change over time, but the prevailing pattern has been a decrease in the gap between students with disabilities and the general student population. In any case, KIRIS does

FIGURE 15-4
Differences in Student Performance Between Students With Disabilities and General Population
Grades 4, 8, and 11, KIRIS Scale Score, 1992-93 through 1995-96



not appear to have disadvantaged students with disabilities, insofar as there have been gains over time and the gap in achievement is closing.

There are also different patterns of services provided to students with disabilities across Grades 4, 8, and 11. Depending on grade level and year of testing, approximately 4% to 10% of the student population have been identified as having disabilities and as having taken the KIRIS assessment with accommodations. (The increase over time in each grade level is consistent with national trends.) It is important to note the percent of students identified as having disabilities declines from elementary to middle school to high school. The percentage of students identified as having disabilities was approximately 10% in Grade 4, 5% in Grade 8, and 3% in Grade 11. The percent of students with disabilities who receive accommodations in taking the KIRIS assessment also declines from Grade 4 (83% average in 1995 and 1996) to Grade 8 (68/%) to Grade 11 (62%). The range of types of accommodations used with the KIRIS assessment also decreases from Grade 4 to Grade 11.

Several hypotheses may be proposed to explain these differences in services provided and student achievement. For instance, one might focus on the ease with which accommodations may be made because of school organization. At the elementary school level, a student is typically in contact with one or two teachers most of the day. In this environment it may be relatively easy to identify and implement appropriate instructional accommodations consistently. At the middle and high school levels, where students are typically engaged with four or more teachers, coordinating and delivering appropriate instructional accommodations may be more difficult. Another hypothesis focuses on the social development of students. As children grow older, they may become more reluctant to acknowledge disabilities and less willing to receive accommodations that "make them different" from their classmates. These are only a few possible reasons.

CONSEQUENCES: DOES NOT HAVE UNINTENDED, NEGATIVE CONSEQUENCES

The challenge in considering unintended consequences is that many are unanticipated as well. Although it is difficult to predict all of the changes that might occur as a result of implementing an assessment program, it is possible to enumerate some undesirable, potential consequences. A potential consequence suggested by critics of whole language and other teaching approaches that lend themselves to teaching real world writing skills is that increased emphasis on writing process might lead to a decline in teaching emphasis, and, in turn, student performance, on writing mechanics. A soon-to-be-released comparison of student responses to on-demand writing prompts from 1992-93 and 1995-96, conducted by KDE's contractor and subcontractor indicates that, at least at grade 4, this potential consequence has not been realized in the least. Student performance did not decline over the four years of the second accountability cycle with regard to capitalization, punctuation, and subject/verb agreement. In fact, performance improved in spelling. This strong evidence against negative consequences can be buttressed by further findings, on secondary analysis, that students' responses in 1995-96 demonstrate increased length of response and complexity of sentence structure relative 1992-93 responses.

Evidence with regard to student mastery of writing mechanics at the higher grade levels is somewhat limited relative to the study of fourth graders, but still favorable. Extensive surveys of students and faculty at Morehead State University conducted by Kathryn Mincey indicate that teachers report a dramatic improvement in college students' understanding of, comfort with and willingness to engage in the writing process over the second accountability cycle. Mincey has noted, "contrary to some popular opinion, we have seen freshmen writers fairly evenly holding their own or improving in basic skills, with many more of us seeing improvement in the past five years than decline."[8]

CONSEQUENCES FOR FUTURE STUDY

Although consideration of the potential consequences of an assessment program could go on almost endlessly in theory, research in practice is necessarily limited to studying questions identified as most important, within the constraints of available resources. Consequences considered above represent some of the most pertinent questions, and evidence collected in response to these questions. KDE, its contractor and subcontractors, and others will further examine consequences of KIRIS, and gather diverse evidence to support or refute the validity of inferences made from KIRIS scores.

---

[8] "Writing Portfolios Benefit College Freshmen, Study Says." Kentucky Department of Education Press Release No. 97-020, April 22, 1997.

Appendix 7-1

Principal's and District Assessment Coordinator's Certification of Proper Test Administration Forms

KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM (KIRIS)
TRANSITIONAL ASSESSMENT
1993-94

PRINCIPAL'S CERTIFICATION OF PROPER TEST ADMINISTRATION
Grade 4

District Name and Code:

School Name and Code:

1. Number of students enrolled in the fourth grade as of April 18, 1994 _____ (1.)

2. Number of students tested _____ (2.)

3. Number of students excluded (See note below.) _____ (3.)

4. Number of students not tested (See note below.) _____ (4.)

5. Total number of used student response booklets _____ (5.)
(Lines 2. + 3. + 4. = 5.) If Line 5. does not equal Line 1.,
please explain why at the bottom of this page.

6. Number of unused student response booklets _____ (6.)

7. Total number of student response booklets being returned to Advanced Systems _____ (7.)
(Lines 5. + 6. = 7.)

8. Total number of student response booklets sent to you _____ (8.)
If Line 7. does not equal Line 8., please explain the discrepancy on the bottom of this page.

NOTE: **You must complete test booklets for all students whether they are tested or not.** Student response booklets for those students who are excluded from testing for special education,* limited-English speaking, or medical reasons in accordance with Department of Education policy should have the appropriate reason for exclusion indicated in the proper section on page 1.

* For this purpose, special education students are those eligible for and participating in the alternate portfolio assessment.

The information provided above is correct to the best of my knowledge. I have also read and initialed the reverse side of this form.

_______________________________ _______________

Principal's Signature Date

Discrepancy Notes:

I certify that:

- the information recorded on page 1 of all student response booklets and this form is correct to the best of my knowledge;
- every effort was made to administer the full assessment to all eligible fourth-grade students;
- the testing was done in accordance with the directions in the *Instruction Manual for Test Administrators* and the *Instruction Manual for Assessment Coordinators*;
- I have checked to see that a student response booklet is assigned to every student enrolled on April 18, 1994, whether he/she was tested or not, and all appropriate sections on page 1 of his/her student response booklet have been properly completed;
- I understand that the KIRIS student assessment test booklets are secure materials and, to the best of my knowledge, have not been reproduced by anyone in my school;
- test booklets were distributed randomly as directed in the administrator's manual;
- to the best of my knowledge, no personnel or student in the school shared answers to test questions or otherwise gave out information that would lead to inappropriate test administration;
- test administrators have read and signed a copy of *Appropriate Assessment Practices* and the copies are maintained on file;
- any authorized photocopying of **student responses to the first four common open-response items** has been done in accordance with procedures provided by the District Assessment Coordinator;
- **all** test-related materials are being returned to Advanced Systems as directed;
- assessment materials revealing assessment content that inadvertently remain in the school after such material was to have been returned to Advanced Systems will be secured and returned immediately upon being found; and
- I shall inform the Department and the local district assessment coordinator immediately if any inappropriate uses of assessment materials are discovered within this school.

____
Principal's Initials

KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM (KIRIS)
TRANSITIONAL ASSESSMENT
1993-94

DISTRICT ASSESSMENT COORDINATOR'S CERTIFICATION OF PROPER TEST ADMINISTRATION
Grade 12

District Name and Code:

I certify that:

- the information recorded on page 1 of all student response booklets and this form is correct to the best of my knowledge;
- every effort was made to administer the full assessment to all eligible twelfth-grade students;
- the testing was done in accordance with the directions in the *Instruction Manual for Test Administrators* and the *Instruction Manual for Assessment Coordinators*;
- I have checked to see that a student response booklet is assigned to every student enrolled on February 7, 1994, whether he/she was tested or not, and all appropriate sections on page 1 of his/her student response booklet have been properly completed;
- I understand that the KIRIS student assessment test booklets are secure materials and, to the best of my knowledge, have not been reproduced by anyone in my district;
- test booklets were distributed randomly as directed in the administrator's manual;
- to the best of my knowledge, no personnel or student in the district shared answers to test questions or otherwise gave out information that would lead to inappropriate test administration;
- test administrators have read and signed a copy of *Appropriate Assessment Practices* and the copies are maintained on file;
- any authorized photocopying of **student responses to the first four common open-ended items** has been done in accordance with procedures provided by the Kentucky Department of Education;
- all test-related materials are being returned to Advanced Systems as directed;
- assessment materials revealing assessment content that inadvertently remain in the school after such material was to have been returned to Advanced Systems will be secured and returned immediately upon being found; and
- I shall inform the Department immediately if any inappropriate uses of assessment materials are discovered within this district.

District Assessment Coordinator's Signature

Date

Appendix 7-2
Summary of Performance Events

1994-95 GRADE 4

M1-F As the Wheel Turns

As a group, the students discuss strategies for measuring the length of a very long strip of paper. They record their strategies on the record sheet and implement the one they think is best. In their individual work, they identify two strategies and evaluate the strategies. They also answer a question relating the size of a wheel to the number of times it turns.

M2-F Good Directions

As a group, the students will discuss strategies for giving directions for making figures on geoboards. They will practice the strategies by taking turns giving and following instructions. Individually, they will write directions for making other geoboard figures.

S1-F How A Boat That?

In this task students will:

- predict which of four boat model shapes will float best empty and which will float best with heavy loads;
- experiment by testing the four shapes in water, first empty, and then with cargo;
- determine which shape can hold the heaviest load without sinking or capsizing; and
- individually, describe with boat shapes would be best for various stated purposes.

S2-F Leaf It To Me

As a group, students will create two different classification systems for a set of artificial leaves. Individually, they will create a classification system for a set of beans, and generalize from that activity to suggest other kinds of objects that could be classified, and how classification is used in everyday life.

SS1-F Pictographs

As a group, the students break into two teams to create an original pictograph story, using symbols from a chart provided. The teams exchange stories and try to translate them. Individually, the students explore whether pictographs or written words work better to tell stories. Finally, they describe what the pictographs tell them about early Native American life.

SS2-F Regional Occupations

The group of students is given 12 cards showing different occupations, and three reference maps. On a large map of the U.S., they place each occupation card in the region where that occupation would be most appropriate. The group fills out a chart, explaining why they placed cards for eight occupations in certain regions. Individually, the students identify appropriate occupations for the southeast region and discuss what features of a region determine the occupations that would be best there.

AH1-F I'm a Yankee Doodle

As a group, students use a keyboard and a diagram of the keys to fill in notes missing from the musical notation for the song "Yankee Doodle." Individually, students explore how different instrumentation affects the impact of a song. They then create new lyrics for the melody of "Twinkle, Twinkle Little Star."

AS2-F Little Red Riding Hood

As a group, students watch a videotaped scene from a performance of "Little Red Riding Hood." They discuss how a director could improve the performance. Individually, students answer questions about how to use costumes and sets to improve a production, and about other elements that contribute to good theater.

AS3-F The Play's the Thing

As a group, students read two short dramatic scenes and choose furniture and props to create a stage set for one of the scenes. Individually, they answer questions about the considerations and process of designing a set.

PL1-F Choosing a Good Friend

The group is given two sets of 15 cards, each showing a personality characteristic. As a group, the students discuss the characteristics of a good friend, and then choose the five characteristics from the cards that they consider most important in a friend. Individually, students write letters explaining why a certain friend deserves a "Good Friend" award, and create a poster about being a good friend.

PL2-F Health Care Workers

As a group, students identify the job responsibilities of a variety of health care workers. Individually, each student selects four health care workers to invite to the school's Stay Healthy Fair, and identifies what each worker could tell the students about staying healthy.

VS1-F **The Choice is Yours**
The group will be given information on three applicants for the position of principal of the school. Students will discuss the information and decide what other information would be needed to help them make the best choice. Individually, students describe characteristics of a good principal and select one of the three applicants, explaining the choice.

1994-95 GRADE 8

M1-F **Coordinate Darts**
As a group, students randomly generate coordinates to see how they fall into a triangle drawn on a grid. They use the coordinate data to estimate the size of the total grid area and the size of the triangle. Individually, students do their own version of the group exercise, and then discuss the reliability of both the group and their own estimates. They then apply that exercise to a real-life model involving estimating the area of a lake on a map.

M2-F **Inscribed Figures**
As a group, the students will use rubber bands to inscribe four regular polygons with a circle marked on a geoboard. The group notes the number of sides on each polygon, then measures the circumference of the circle. Individually, students determine the length of the sides and the perimeter of each polygon, producing a graph to show the relationship between the number of sides and the perimeter of the polygons. They then answer questions about estimating perimeters and circumferences, inscribing pentagons, and estimating a circle's area.

S2-F **Where the Rubber Meets the Road**
Students will test the force necessary to start each of two blocks in motion on an asphalt strip. They will repeat their tests with weight on the blocks, and with three different asphalt strips: one dry, one wet and one oily. Individually, they will graph the group's data, and apply the data to several different applications.

SS1-F **Production and Consumption**
As a group, students will list the jobs that are involved in producing a taco served at a fast food restaurant. Individually, students will describe strategies that a fictional company could use to become more competitive and earn more profit.

SS2-F Amendments
As a group, the students will look over the Amendments to the Constitution of the United States of America and discuss how the amendments function in today's society. Individually, students will present their opinions on how the amendments affect their own lives and which one is the most important of the amendments today.

AH1-F A Different Dimension
As a group, students will analyze and discuss Da Vinci's painting "The Last Supper." Individually, they will draw a picture showing one-point perspective, and describe how a horizon line and a vanishing point create the illusion of three-dimensionality in two-dimensional art.

AH2-F Money Makes the World Go 'Round
As a group, students will discuss elements of the design of the U.S. one dollar bill. Individually, they will draft a design for a new dollar, and describe their reasons for choosing their new designs.

AH3-F Won't You Play a Simple Melody?
As a group, student will take turns playing the melody of "When the Saints Go Marching In' on the keyboard. Then they will listen to the song "Au Clair de la Lune" on the audiocassette provided. Individually, they will compare the two melodies and analyze the melodic contour of "When the Saints Go Marching In."

PL1-F Fun and Fitness
As a group, students determine whether teenagers participate in listed activities for fun, physical fitness, or both. They explain the reasons for their choices. Individually, each student chooses an activity. In a series of questions, students describe the benefits derived from the activity, skills needed to participate successfully in that activity, and other factors that could contribute to success in the activity.

PL2-F Shopping Spree
As a group, students review the product displays and cost information on the Information Sheet. They discuss and record the factors that would contribute to their decision whether or not to buy each product. Individually, they describe how they would advise a friend to allocate money on these products, and then describe how they would each spend $25.00.

VS1-F On the Job

As a group, students select seven of the occupations prohibited to minors according to the Kentucky Child Labor Law. They discuss reasons why those occupations are prohibited for young workers and identify the health/safety hazards that may be involved. Individually, they identify a variety of occupations, listing some positive and negative aspects for each, and explain the factors a person should consider when making a career choice.

1994-95 GRADE 11

M1-F Logo

As a group, the students discuss how to determine the height of three equilateral triangles used in a logo, and how to divide one large triangle into three smaller triangular regions. Individually, the students will use that information to determine the height the logo needs to be to fit a given background size.

M2-F Traffic Flow

Students use matrices of information about traffic flow through an intersection and the traffic light cycle at that intersection to organize data and make calculations about traffic flow and cycles of traffic lights. Individually, the students use the information to create additional matrices and draw further conclusions about the traffic flow to determine if that flow is efficient.

S1-F Let the Sunshine In

Students set up equipment according to a diagram and perform an experiment by recording the temperature of three different-colored cups every 30 seconds for eight minutes. Individually, they organize and analyze the data, using it to make a recommendation on the design of a sun-filtering product.

S2-F Classification in Action

As a group, students follow an example to design and complete two classification systems using a set of objects provided. Individually, they describe how another object would fit into one of their systems, describe other classification systems the group could have used, and relate classification to everyday life.

SS1-F Historical Perspective

The group examines examples of artifacts of the Gilded Age in American history and chooses those they think best exemplify that age. Individually, they analyze the impact of one artifact and describe a contemporary counterpart.

SS2-F Propaganda

Students examine political cartoons and discuss the propaganda techniques that make them effective. Individually, they describe the role of propaganda in modern life and design their own visual propaganda piece.

AH1-F Beauty in the Eye of the Beholder

Students in the group examine a set of photos of works of art, discussing the artists' perceptions and the elements and principles of art present in the works. Individually, they compare the artists' use of elements and principles, and describe how their own definition of beauty support a choice of the "most beautiful" of the collected works.

AH2-F *Death of a Salesman*

The group analyzes the script of a scene from *Death of a Salesman* to discuss stage sets and stage directions for the excerpt. Individually, students reinterpret the scene to alter its mood, set, time period, or environment, and then describe different physical attributes that could be used to stage their versions.

AH3-F Special Effects

The students in the group watch a videotape showing clips from various films. They discuss how the special effects in the clips might have been created and how they added to the scenes. Individually, they explore the differences between special effects in film and those used in live theater. They also analyze more comprehensive differences between film and live theater.

PL1-F Play on

As a group, students act as members of the planning committee for a new community recreation center. They discuss the kinds of physical activities the center could offer to stimulate participation of people from various age groups. Individually, students select two physical activities they feel would encourage lifetime involvement for all age groups. In a letter to the center's board of directors, students explain the specific health and fitness benefits of both activities.

VS1-F **Candid Camera**

As a group, students watch a video of a young woman being interviewed before and then after some coaching. They discuss what the young woman did well and needed to improve, both "before" and "after" the coaching. Individually, students choose a position (job) that interests them and describe how they would prepare for an interview for this position. Finally, they write a script of questions they might be asked during the interview and how they would answer.

1995-96 GRADE 4

M1 **Different Places, Different Pictures**

As a group, the students place models of a tree, a tank, and a rock on a map. Presented with several drawings of the objects, they determine where on the map a person would have to be standing to see the objects in the positions shown. Then they draw how the objects would look from other locations. Individually, they repeat the tasks using a different map.

M2 **Run, Dog, Run**

As a group, the students create different arrangements of five tiles shaped like rhombuses. They draw the arrangements that are not congruent to one another. Individually, students find different arrangements of four tiles, again drawing the arrangements that are not congruent. They explain how to determine whether arrangements are congruent.

S1 **Pathfinder**

As a group, students will use a battery pack and bulb to study an electrical circuit card. The card is open underneath so some connections between the contacts are exposed. Then they will observe surface connections on a closed card to locate the hidden connections that make the bulb light. Individually, students will describe strategies that they think are best for investigating the structure of the hidden connections in the circuit card and in other electrical systems.

S2 **Temperature School**

Students will work with three cans of water: one at room temperature, one hot, and one ice cold. They will observe the cooling and warming of the hot and cold cans by measuring the change of temperature over time and comparing the measurements to those of the room temperature water. They will record their measurements in a table. Individually, they will explain the observations made by the group and apply their findings to actual situations.

SS1 Advertisements

As a group, students decide what pieces of information are required for an ad to be fully informative. They examine four advertisements to see what information is missing. Individually, they list the information a good ad needs to include, and design an ad for a class car wash.

SS2 Democratic Principles

As a group, students assign symbols to the principles they think best represent: freedom, responsibility, equality, and justice. Individually, students choose an image to represent one of the principles or design their own and explain their choice or design. They then explain why the principle they each chose is beneficial to life in the United States.

AH1 Artist

As a group, the students look at two famous works of art and discuss the artists' uses of warm or cool colors and explains the choice of color scheme and the mood or feeling the colors were intended to convey.

AH2 Clay Works

As a group, the students look at pictures of clay pots, discuss the pots, and brainstorm ideas for other useful clay items. Individually, each student creates a cup, bow, plate, or vase out of clay.

AH3 Musical Composition

As a group, the students work to play "Mary Had a Little Lamb" from printed notation. The group creates a new melody to play and discusses its use of steps, leaps, and repeated tones. Individually, students listen to two melodies and describe how each one moves, then describe how the melodies are alike and different.

PL1 Cereal Selection

As a group, the students discuss important things people should think about when choosing a cereal. They then examine three different cereal boxes and discuss and list the words and pictures used to persuade people to buy the cereal. Individually, students each design a box for a new children's cereal; the instructions specify that they are to include nutritional information in their designs.

PL2 It's a Toss Up

The group will watch a videotape showing examples of the correct way to throw a ball underhand, and will discuss the individual steps that make up a correct underhand throw. Individually, each student will open an envelope containing six pictures, each showing one step in the underhand throw technique. The students will list the correct sequence of the pictures, and will describe how they would explain these steps to a younger student.

VS1 Work Webs

The group will examine the skills listed in the center of the work webs on the back of the response form and decide which jobs need those skills, listing their choices in the circles on the web. Individually, students will choose a job and list on a work web the skills needed to do that job. They will also describe how they would explain to younger students that several of their school subjects would help someone perform that job.

1995-96 GRADE 8

M1 Simulation

Students label and toss a 10-sided polyhedra to simulate the probability of getting a complete set of game pieces one at a time from boxes of cereal. They compute the mean, median and mode of the results of the polyhedra tosses and discuss how these statistics indicate how many boxes of cereal they would need.

M2 Stained Glass Window

Students determine how many of each color and shape pattern blocks to use to fill a shape according to instructions that specify the percentage of each color.

S1 Effect of Temperature on Enzymes

Students will follow a procedure to add rennin to both cold and room temperature mile, observing the results. They then analyze the experiment and its results.

S2 Half-Ball

Students follow a procedure to measure the bounce of a racquetball and a half-ball. They relate the results to Newton's Laws of Motion, producing a graph and describing conclusions based on the data.

SS1 The Candidate

The group creates a platform for a presidential candidate, outlining the candidate's position on four national issues. Individual students discuss the importance of two of those issues for Kentucky. They also identify an additional issue important to Kentucky and state a position on that issue.

SS2 Sequencing

The group determines the sequence of a set of eight pictured events. Individually, the students describe the impact of one of the historical events and explain how cause and effect apply to that event.

AH1 Costume Design and Character Creation

The group reads a scene from *Our Town* and discusses what the characters' costumes would look like. Individually, students design a costume and explain why they made their choices of design and fabrics.

AH2 Sales Pitch

The group listens to a musical selection and lists the instruments heard. Individually, students describe two band instruments in detail, and then select two instruments not commonly found in a band, comparing them to band instruments.

AH3 Traveling Exhibit

The group discusses reasons for including certain works of art in a traveling exhibit for elementary students. Individually, students analyze and describe one artwork to create an information sheet for an exhibit.

PL1 Health Habits

The group will empty the contents of a "grab bag" containing personal care items and discuss how each item is used and related to health. Individually, students develop a public service announcement promoting good habits for health and hygiene.

PL2 Safety Fair

The group will list and discuss topics that would need to be included in a community safety fair. Individually, students list topics, name agencies or groups that could provide speakers on each topic, and explain why they chose the agencies or groups. They will also create a flyer advertising the safety fair.

VS1 Come to Order

Each student selects a card naming a role to play. The group role plays a discussion about an appropriate school reward for perfect attendance. After five minutes they discuss the role play. Individually, students list rules or guidelines that could ensure that meetings are well-run and productive and explain how the guidelines would work.

1995-96 GRADE 11

M1 Box and Whisker, Stem and Leaf

The group discusses the use of both kinds of plots and computes statistics for given sets of data. Individually, students further analyze the use of such plots and draw their own conclusions for new data.

M2 Miniature Golf

The group rolls a ball toward a solid surface and measures the angles created by its path. They discuss possible paths for a hole-in-one on a particular miniature golf hole. Individually, they predict the path necessary for a hole-in-one at a different miniature golf hole, and design their own hole according to certain guidelines.

S1 A Cut Above the Rest

The group conducts an experiment to measure both the vertical force required to move a scissors blade and the vertical distance the blade moves, and calculates the output work. Individually, the students graph and interpret the data from the experiment, and analyze the results in terms of the Law of Conservation of Energy. Students then design scissors suitable for a small child.

S2 Coat Caper

The group plans and conducts a procedure to compare samples of pollens and fiber from a garment with slides of pollens and fiber consistent with a historical period. The group records its data. Individually, students display the group's data in a chart and write a letter documenting their conclusions regarding the garment's authenticity for that time period.

SS1 The Big Dig

As a group, students study a map and record their observations about what the map tells them. Individually, they use the observations to draw conclusions about the culture identified on the map. Next they draw their own maps showing what an archaeologist might find in an American area several thousand years from now.

SS2 Presidential Speech

Students watch a dramatization of George Washington delivering a speech. They discuss the ideas in the speech about political parties, and discuss positive and negative aspects of political parties of today. Individually, they state their own agreement or disagreement with Washington's ideas, and describe what they think the role of political parties should be in the U.S. today.

AH1 Classical Theater and Commercials

The group reads a short excerpt form *Romeo and Juliet* and discusses the concepts Shakespeare conveyed. Individually, the students adapt one of the ideas from the group discussion for use in an advertising campaign including a television commercial and a print advertisement.

AH2 Dance and Culture

The group of students watches a videotape showing dances from four different cultures. The students discuss the styles and forms of the dances, and matches each dance with a culture they think it comes from. Individually, students choose two of the dances and compare their styles and forms, then describe and compare their purposes and uses.

AH3 Great Painters

As a group, students discuss similarities and differences in the styles of portraits painted by four artists. Individually, students paint portraits, each incorporating the style of one of the four artists, and explain the aspects of the style they tried to use.

PL1 Design a Community Health/Safety Center

As a group, students discuss and list all the health and safety groups within a community. Individually, they select three to four agencies or groups that they feel are very important and should be located in a new Community Health/Safety Center, and explain why they think they are important.

PL2 Peer Pressure at School

As a group, students will read and discuss three letters by students to an advice column addressing social issues. They list the similarities and differences among the situations described in the letters. Individually, students draft a school newspaper article addressing all of the letters and outlining positive strategies for dealing with the situations.

VS1 Résumé Review

As a group, students review and discuss four different résumés and list their strengths, weaknesses and strategies. Individually, they create draft résumés for themselves and explain why they chose to organize them as they did.

Appendix 9-1
Accountability Cycle II Equating Plan

EQUATING PLAN: ACCOUNTABILITY CYCLE II

OPEN RESPONSE TESTS IN
READING, MATHEMATICS, SCIENCE, AND SOCIAL STUDIES

ITEM CALIBRATION

1. Calibrate all 1992-93 items within a subject area in a single Multilog run. That is, all 12 forms of a test will be calibrated using the common items to link the theta metrics.

   A. Remove from the sample any examinees meeting one or more of the following criteria: a total raw score of 0, exclusion from accountability testing, indications of improper administration such as taking multiple forms, evidence of lack of attempt to respond seriously to any items. Document how many examinees fell into each category.

   B. Document sample characteristics with common item raw score distributions , mean and standard deviation for each form sample.

   C. Document calibration results with separate frequency distributions, means, and standard deviations for the a, b_1, b_2, b_3, and b_4 parameter estimates. Include the Multilog output as an attachment.

2. Calibrate all 1993-94 items within a subject area in a single Multilog run. That is, all 12 forms of a test will be calibrated using the common items to link the theta metrics.

   A. Remove from the sample any examinees meeting one or more of the following criteria: a total raw score of 0, exclusion from accountability testing, indications of improper administration such as taking multiple forms, evidence of lack of attempt to respond seriously to any items. Document how many examinees fell into each category.

   B. Document sample characteristics with common item raw score distributions , mean and standard deviation for each form sample.

   C. Document calibration results with a frequency distribution, mean and standard deviation for the a, b_1, b_2, b_3, and b_4 parameter estimates. Include the Multilog output as an attachment.

3. Calibrate all 1994-95 items within a subject area in a single Multilog run. That is, all 12 forms of a test will be calibrated using the common items to link the theta metrics.

    A. Remove from the sample any examinees meeting one or more of the following criteria: a total raw score of 0, exclusion from accountability testing, indications of improper administration such as taking multiple forms, evidence of lack of attempt to respond seriously to any items. Document how many examinees fell into each category.

    B. Document sample characteristics with common item raw score distributions , mean and standard deviation for each form sample.

    C. Document calibration results with a frequency distribution, mean and standard deviation for the a, b_1, b_2, b_3, and b_4 parameter estimates. Include the Multilog output as an attachment.

4. Calibrate all 1995-96 items within a subject area in a single Multilog run. That is, all 12 forms of a test will be calibrated using the common items to link the theta metrics.

    A. Remove from the sample any examinees meeting one or more of the following criteria: a total raw score of 0, exclusion from accountability testing, indications of improper administration such as taking multiple forms, evidence of lack of attempt to respond seriously to any items. Document how many examinees fell into each category.

    B. Document sample characteristics with common item raw score distributions , mean and standard deviation for each form sample.

    C. Document calibration results with a frequency distribution, mean and standard deviation for the a, b_1, b_2, b_3, and b_4 parameter estimates. Include the Multilog output as an attachment.

RESCORING ANALYSIS

5. Conduct a rescore analysis for 1992-93 and 1993-94 to ensure scoring standards were applied the same way each year.

   A. Score about 200 1992-93 responses (50 each 1,2,3,4) for each equating common item during the 1993-94 scoring.

   B. Since regression effects might lead the rescores of 2, 3 and 4 responses to be lower than the original scores (and the rescores of 1 responses to be somewhat higher)[1], recast the results to minimize this effect.[2]

   C. Compare the recast mean of the 1993-94 scores on the set of papers with the recast mean of the 1992-93 scores on the same set of papers. If a substantial difference is observed in the mean scores across items, examine recast mean differences for individual items to determine which should be eliminated from the equating.[3]

   D. Document the rescore analysis for review by the National Technical Working Group.

6. Conduct a rescore analysis for 1993-94 and 1994-95 to ensure scoring standards were applied the same way each year.

   A. Score about 200 1993-94 responses (50 each 1,2,3,4) for each equating common item during the 1994-95 scoring.

---

[1] Since most items have means between 1.5 and 1.9.

[2] An alternative method (perhaps easier to explain) might be implemented in the future. Choose the 200 item responses from the approximately 1,000 responses that were double scored. Compare the mean of the rescores by the year 1 table leaders with the mean of the rescores by the year 2 scorers. Any item for which the means are significantly different at the .05 level are not considered common items for equating purposes, unless the total number of equating points falls below 40. If the number of equating points falls below 40, discard only items for which the difference between means is significant at the .01 level.

[3] If it becomes necessary to eliminate individual items from the equating, one could reasonably argue that a decision rule for doing this should consider effect size in addition to or in lieu of statistical significance. In fact, there are a number of complex factors that will affect the decision to include or exclude an item from the common equating item set. This step should be taken as a general guideline, not a hard and fast rule.

B. Since regression effects will lead the rescores of 2, 3 and 4 responses to be lower than the original scores (and the rescores of 1 responses to be somewhat higher), recast the results to minimize this effect.

C. If a substantial difference is observed in the mean scores across items, examine recast mean differences for individual items to determine which should be eliminated from the equating.

D. Document the rescore analysis for review by the National Technical Working Group.

7. Conduct a rescore analysis for 1994-95 and 1995-96 to ensure scoring standards were applied the same way each year.

A. Score about 200 1994-95 responses (50 each 1,2,3,4) for each equating common item during the 1995-96 scoring.

B. Since regression effects will lead the rescores of 2, 3 and 4 responses to be lower than the original scores (and the rescores of 1 responses to be somewhat higher), recast the results to minimize this effect.

C. If a substantial difference is observed in the mean scores across items, examine recast mean differences for individual items to determine which should be eliminated from the equating.

D. Document the rescore analysis for review by the National Technical Working Group.

LINKING THETA METRICS

8. Place the 1993-94 parameter estimates on the 1992-93 theta metric (scale) by setting means and standard deviations equal in an iterative process.

A. For both the 1992-93 and 1993-94 estimates, calculate the mean and standard deviation of the b-parameter estimates of the items that appeared in the matrix section of both the 1992-93 and 1993-94 test. Because the definition of score points of 0 and 1 were changed in 1993-94, include only b's corresponding to the cutpoints for 1-2, 2-3, and 3-4.

B. Calculate the parameters for a linear equating of theta metrics, based on the full set of b-parameter estimates as described in the preceding step:

$$\text{slope}_{93\text{-}94 \text{ to } 92\text{-}93} = \quad (\sigma_{b\ 92\text{-}93})/(\sigma_{b\ 93\text{-}94})$$

$$\text{intercept}_{93\text{-}94 \text{ to } 92\text{-}93} = \quad \mu_{b\ 92\text{-}93} - (\mu_{b\ 93\text{-}94})(\text{slope}_{93\text{-}94 \text{ to } 92\text{-}93})$$

C. Calculate an equated set of b's for each 1993-94 item used in step A:

$$b_{93\text{-}94 \text{ new}} = (\text{slope}_{93\text{-}94 \text{ to } 92\text{-}93})(b_{93\text{-}94 \text{ old}}) + \text{intercept}_{93\text{-}94 \text{ to } 92\text{-}93}$$

D. Calculate $\Delta_b = b_{93\text{-}94 \text{ new}} - b_{92\text{-}93}$

E. Remove every item for which a pair of b's has $|\Delta_b| \geq .4$. If less than 40 b's remain after doing this, replace all items and proceed to step F.

F. Remove every pair of b's for which $|\Delta_b| \geq .4$. If the resulting number of b's is 32 or greater, proceed to step H. Otherwise, proceed to step G.

G. If fewer than 32 b's remain based on the matrix items only, repeat steps A-F using matrix item b's and common item b's. If fewer than 32 b's remain based on the matrix and common items, retain the 32 b's having the lowest $|\Delta_b|$ and proceed to step H.

H. Repeat steps A-B based on the selected b's.

I. Apply the final equating transformation from step H to the new form (1993-94) theta estimates for each student:

$$\text{THETA}_{93\text{-}94\text{new}} = (\text{slope}_{93\text{-}94 \text{ to } 92\text{-}93})(\text{THETA}_{93\text{-}94\text{old}}) + \text{intercept}_{93\text{-}94 \text{ to } 92\text{-}93}$$

9. Repeat steps 8A through 8I to find and apply the slope and intercept to equate the 1994-95 theta estimates to the 1992-93 metric via the 1993-94 results. That is, substitute transformed 1993-94 item parameter estimates in place of 1992-93 raw item parameter estimates and 1994-95 raw item parameter estimates for 1993-94 raw item parameter estimates in the carrying out steps 8A through 8I. The transformation for the 1993-94 item parameter estimates was found above, and is the same transformation as was applied to theta estimates:

$$b_{93\text{-}94\text{new}} = (\text{slope}_{93\text{-}94 \text{ to } 92\text{-}93})(b_{93\text{-}94\text{old}}) + \text{intercept}_{93\text{-}94 \text{ to } 92\text{-}93}$$

Using transformed 1993-94 items parameters in this manner yields linkage of 1994-95 scores to the 1992-93 scale rather than the 1993-94 scale.

10. Repeat steps 8A through 8I to find and apply the slope and intercept to equate the 1995-96 theta estimates to the 1992-93 metric via the 1994-95 results. That is,

substitute transformed 1994-95 item parameter estimates in place of 1992-93 raw item parameter estimates and 1995-96 raw item parameter estimates for 1993-94 raw item parameter estimates in repeating steps 8A through 8I. The transformation for the 1994-95 item parameter estimates was found above, and is the same transformation as was applied to theta estimates:

$$b_{94\text{-}95new} = (slope_{94\text{-}95 \text{ to } 92\text{-}93})(b_{94\text{-}95old}) + intercept_{94\text{-}95 \text{ to } 92\text{-}93}$$

Using transformed 1994-95 items parameters in this manner yields linkage of 1995-96 scores to the 1992-93 scale rather than the 1994-95 scale.

11. Document the equating, including (1) characteristics of the items (identification of form and location of equating common items, items removed from being considered common, distributions, means and standard deviations of parameters for common items on both samples, slope and intercept to place b parameters and thetas on the 1992-93 theta metric) and (2) characteristics of the equating samples (sample size, number of excluded examinees, and distribution of transformed thetas for sample).

NTWG REVIEW

12. Meet with the National Technical Working Group to review all equating-related decisions and equating documentation. Consult with National Technical Working Group members individually when it is necessary to review equating-related decisions and equating documentation more promptly than scheduled meeting times allow.

Appendix 9-2
Adjustments for Changes in Scorer Leniency

APPENDIX 9-2

ADJUSTMENTS FOR CHANGES IN SCORER LENIENCY FROM 1994-95 TO 1995-96

This appendix illustrates procedures used to determine whether scoring standards changed enough from 1994-95 to 1995-96 to merit an adjustment, and, if so, the cases in which such an adjustment was appropriate. The procedure developed to implement an adjustment for systematic changes in scoring standard application (which differs from the approach called for and the difficulty anticipated in the equating plan) is illustrated, along with some evidence that the approach implemented was reasonable. Grade 11 rescoring is presented first, followed by the more complex procedures required for grades 4 and 8.

DETERMINING WHETHER SCORING STANDARDS CHANGED SIGNIFICANTLY

Year-to-year shifts in the application of scoring standards can be detected by rescoring a sample of papers from a base year, and comparing in some manner the original mean of the rescored papers with the mean obtained after rescoring. The most appropriate comparison is between the mean of all base year papers (not just those that were rescored) with the expected mean if all had been rescored. The expected mean can be obtained by "recasting" the distribution of the rescored papers against the stratification used in sampling, which means weighting each rescore by the relative frequency of the original score assigned to each paper. If twenty papers were drawn to represent each of four score points assigned with the respective frequencies .1, .2, .3 and .4, then rescores from the first set of twenty papers would be weighted .1 in computing a recast mean from which to infer the expected value if all such papers had been rescored.

Recast means of 1992-93 papers rescored in 1993-94 and 1993-94 papers rescored in 1994-95 were computed based on the marginal distribution of item scores, as decribed above; papers were sampled with stratification on item responses, 1 through 4, and recast means were computed by taking the percentage of students receiving each of these responses, and multiplying the percentage by the means obtained on papers originally scored at the given point. In making these comparisons, a recast mean difference as large as .10 was not observed when a stratified sample of papers from one year were scored a second time with papers from the following year. This same procedure was followed for grade 11, rescoring 1994-95 papers during 1995-96 scoring, with the unusual result that differences in magnitude of about .10 were observed in three content areas: Social Studies, Arts & Humanities (AH), and Practical Living/Vocational Studies (PLVS). Advice KDE obtained from its National Technical Working Group (NTWG) at its August, 1996, meeting indicated that it was acceptable to take corrective action in these cases, based on never having found such a difference, and that it may be reasonable to do so in cases where a recast mean difference as large as .05 occurs. KDE elaborated upon this approach by implementing a statistical test of significance to support the use of this criterion, in that it is not desirable to implement changes when recast mean differences fall less than two standard errors from zero. (Because statistical tests of significance will yield significant results for even an inconsequential difference as long as a sufficiently large sample is used in estimating the differences, such tests are probably not the best sole indicator of whether to apply or not to apply a particular difference; they are, however, a good indicator of whether a given difference is not clearly not due to chance, and therefore worthy of consideration for applying a correction. Expert judgment is a better basis for determining whether a difference is of sufficient magnitude to necessitate an adjustment.)

To compute a standard error, it was necessary to consider that the variance of the recast mean estimate is the variance of a composite score, i.e., the sum of the variance of the mean estimates (for papers sampled as score point 1, score point 2, score point 3, score point 4 on each item) multiplied by the square of their weights. The standard error of a recast mean score is equal to the square root of the summation across score points of: the weight (squared) times variance of sampled student scores, divided by the number of students in the random sample. Weight in this case is the percent of non-zero papers each score point represents. (Note that a finite population correction may be applied as well.) As shown below, at grade 11, recast mean differences greater than .05 were more than two standard deviations from zero. Note that the raw score mean differences correspond in magnitude, across content areas, with the log-odds based adjustments implemented (described in the next section), i.e., the larger the recast mean score difference, the larger the log-odds based adjustment applied.

Grade 11 Rescore Statistics

Content Area	Recast Raw Mean Score Difference*	Standard Error of Difference	Log-odds Adjustment Applied
Reading	.014	.020	(none)
Mathematics	.044	.016	(none)
Science	-.021	.021	(none)
Social Studies	.092	.017	.29
Arts & Humanities	-.118	.030	-.31
Practical Living/ Vocational Studies	-.158	.032	-.46

*negative values indicate that scoring became more lenient

IMPLEMENTING AN APPROPRIATE ADJUSTMENT AT GRADE 11

With grade 11 Social Studies, a case discussed in detail at the August, 1996, NTWG meeting, the difference in scoring standards was thought to be appreciable and systematic. That is, a recast mean difference of about .10 was found on average, which is substantial as judged against the standard of previous years' rescoring as well as a two-standard-errors-different-from-zero standard, and is the result of several items having differences substantially exceeding .10, rather than an aberrant result for one or two of the twelve rescored items. This pattern of differences suggests that dropping items (although this was called for in the equating plan) was not the most desirable strategy for making an equating adjustment. The fact that dropping the three most harshly-rescored items did not appreciably change the equating suggested that the item elimination strategy was problematic. The additional finding that the equating based on dropping these three items yielded statewide accountability index scores indicative of a precipitous drop in student Social Studies performance from 1995 to 1996, compared with essentially no changes in performance in Reading, Math, and Science for the same years, corroborates to some degree the notion that dropping items is not satisfactory. Eliminating a larger number of items from the equating pool based on their having been rescored too differently from the previous year would reduce the number of item well below the desired number of items for equating. Therefore, a new approach was developed.

When systematic scoring changes occur, a global adjustment needs to be made, or scoring needs to be done over again. The challenge of making a global adjustment is that it is difficult to calculate the operational impact of systematic changes to scoring standards on estimation of item parameters, the basis of the equating. KDE and ASME jointly conducted analyses of the grade 11 data to test available methods.

EXPLORATION OF LINEAR ADJUSTMENT. The first family of adjustments considered by KDE/ASME involved corrections to the mean and standard deviations of item parameters. Two strategies were initially noted at the NTWG meeting as possible methods to obtain global score adjustments, given the time and personnel resources available for this activity.

A. Re-estimate item parameters for the rescored items holding original (1995) theta estimates fixed.

B. Re-estimate item parameters for the rescored items treating them as unique items in a calibration run.

Following correspondance with the author of Multilog, the feasibility of these two strategies was reconsidered. The problem with method A, as pointed out by David Thissen, is that "theta estimates are biased (shrunken) inward . . . so estimates of item parameters computed by regressing the item responses on those thetas will be biased (b's inward, slopes upward). [There is] Not a lot of effect the first time you do it, but it may accumulate. It is not a well-studied problem" (personal communication). Also, this method does not yield estimates of standard errors of estimation for item parameters. That is, Multilog output provides no indication of the parameter standard error in this case.

For method B, standard errors of item parameter re-estimates based on eighty students per item used in re-estimating parameters were large, ranging from .15 to 1.42. This was not unexpected, given the small number of students used to calculate item parameter estimates, but was appreciably higher than the median of .05 for matrix-sampled item parameters estimated under ordinary conditions. Another potential pitfall of this method is the lack of linkage to the original scoring. That is, student responses scored by one set of scorers are calibrated with student responses scored by a different set of scorers who are believed (in advance of the analysis) to have applied a systematically different scoring standard.

Based on the need to develop an operationally feasible method that could also be applied to AH and PLVS, which are scored using log-odds ratios rather than relying upon Multilog, Advanced Systems developed an alternative approach. This approach (label it 'C'), is based on the comparison of log-odds ratios (computed in essentially the same manner as used for AH and PLVS, only based on recast percentages rather than results for the score points alone) from original scores and rescores. The adjustment is the linear transformation that would need to be applied to the log-odds ratios computed from the rescored items to place them on the same scale as log-odds ratios computed from the original scores assigned to those items.

Finally, as a check on Approach C, KDE developed a probability-based rescoring approach. This approach (D) employed random reassignment of item-level scores based on empirical score/rescore probabilities, followed by re-estimation of item parameters using the reassigned item scores (described below). Of course, this method confounds changes in scoring standards with expected regression to the mean due to rescoring, and introduces error into the rescoring based on randomness in score reassignment. Method D is more of a check on other methods than an operationally feasible approach.

The following table compares results obtained for grade 11 Social Studies under these methods, both in terms of the slope and intercept of the obtained linear correction and the resulting KIRIS index. Item parameters from methods A and B correlate .998, with A indicating some shrinkage (as noted by Thissen; in rough terms, about ten percent). For simplicity, results from B alone are presented below.

Scorer Correction--Grade 11 Social Studies

Method	Slope/Intercept	Mean KIRIS index
B	0.8316/-.2341	35.2
C	0.9155/-.1671	32.1
D	1.0296/-.1134	29.2

One concern with method B is that item parameter re-estimates based on rescoring correlate with original parameter estimates only about .87. Eliminating several re-estimated item parameters with standard errors greater than .31 brings this correlation up to .94, which is still less than the correlation between the recast log-odds based item parameters of method C and Multilog's item parameters (.96, without elimination of any outliers). Following this result and further consultation with Robert Mislevy, it was decided that method B was problematic when based on the small numbers of rescored papers, but also that Method D did not provide as adequate a check on the process as would be desired. It was also apparent that further results should be examined primarily with regard to the theta scale than the KIRIS performance level index reported above.

ADDITIVE ADJUSTMENTS, NOT LINEAR. After further consultation with NTWG members and Robert Mislevy, KDE and ASME concluded that the log-odds based approach was reasonable, given the timeline and operational constraints of score release, and that Multilog held little promise in a situation where item parameters would have had to be estimated from a somewhat limited number of observations per item. The second conclusion was that the adjustment should be additive rather than linear. This was based less on any particular advice provided by NTWG members than on a reconsideration of should be expected from samples of the size employed. Given modest sample sizes of 80 students per item for rescored papers, and a total of seven to thirteen rescored items, estimation of a mean shift is a reasonable demand to place on the data. More subjects and items would be desirable, particularly for estimation of changes in standard deviation. Grade 11 data in this regard might better support a linear correction than the grade 4 and 8 data described later; a sample size of about 500 is probably required to be reasonably certain that an accurate fix has been implemented relative to the error, let alone bias, introduced by any such manipulation of the data. Concern over bias is particularly important in this case, as rescoring items on a four point scale typically results in an increased variability in log-odds difficulties as a result of regression to the mean. Although simplicity of explanation is secondary in importance relative to technical considerations, an additive adjustment, which effectively augments each students' score by the same amount, is more readily explained to the broader public in terms of compensating each student equally for stringent or lenient scoring.

A SIMULATION TO ESTIMATE UPPER AND LOWER BOUNDS OF SCORING EFFECTS. Given that equating is based on item parameters, it is useful to illustrate the impact of differentially harsh scoring on the item parameters obtained from a given set of papers comprising the open-response portion of KIRIS. Item parameters were estimated once using observed grade 11 responses from spring 1996. These values were used for equating back to 1995 item parameters. However, 1995 responses that were scored a second time with the 1996 responses indicated that relatively harsh scoring standards were applied in 1996. The following analysis was conducted to help answer the question, what 1996 item parameters would have been obtained under scoring standards more explicitly comparable to those of 1995 (i.e., easier standards than were apparently applied in 1996):

1. Generate a probability distribution describing the relationship of 1996 scores to 1995 standards, i.e., the conditional probability that an observed 1996 score would turn out to be a given 1995 score.
2. Apply that probability distribution to all 1996 responses to yield probabilistically altered score vectors for all students. For some students, no changes were made. Overall, changes resulted in higher scores at the item level, due to easier standards of rescoring.
3. Estimate item parameters based on students' probabilistically altered score vectors, which reflect less harsh scoring standards than responses used in the original estimation.
4. Compare item parameters from step 3 with the original item parameters.

To address the question of what transformation should be applied to place 1996 item parameters and student theta scores on a scale reflecting the same degree of scorer harshness as the 1995 scores, 1996 item parameters obtained under conditions simulating less harsh scoring were compared with 1996 item parameters obtained using the observed student responses, and were used to score the probabilistically rescored vector of responses for each student and the original vector of responses for each student. Probabilistic rescoring yielded item parameters with mean .825 and standard deviation 2.082, compared with mean .920 and standard deviation 2.002 for the original responses. (These comparisons are based on three thresholds per item and twelve items used in the equating, based on the four score points per item that were rescored). A mean difference of .095 is a rough indication of the degree to which the test is made easier when the more lenient scoring standard is applied. (An alternative way of expressing this differences is to note that a linear transformation with slope 1.0401 and intercept negative .1320 could be applied to the original 1996 item parameters to place them on the scale of the easier scoring standard, and that the two sets of item parameters correlate .997.)

Mean theta is roughly equivalent when the original vector of item responses is scored using item parameters from the original file (mean=.00) and when the probabilistically rescored responses are scored using the item parameters obtained under conditions simulating less harsh scoring (mean = .02). However, when the original vector of item responses is scored using the item parameters obtained under conditions simulating less harsh scoring, mean = -.07. This is sensible, as lower theta scores are expected for the same vector of item responses when item parameters indicative of an easier test are applied. Another way of considering this result is that a mean of shift of -.07 theta occurs when lower item parameters are applied to a given set of responses. Note, however, that the mean of the item parameters is .095 lower in this case, relative to a mean shift of .07 theta. This difference is due in part to the introduction of additional error into the scores through the probabilistic rescoring approach. Larger error leads to less extreme theta scores under Bayesian estimation. (A relevant distinction is that item parameters are estimated using a maximum likelihood approach, whereas student theta estimates result from a Bayesian approach. Student ability estimates will roughly fit a normal (0,1) distribution under any scoring condition, whereas the item difficulty parameters will be higher under conditions of harsher scoring and lower under conditions of more lenient scoring.)

To address considerations of error introduced into rescoring, a second set of probabilistically rescored responses was generated using the approach of assigning higher scores (easier rescoring) not merely at random, but conditionally, to students at each score point having the highest raw score. This approach attempted to model the extent to which not all 2's are equally '2'; some could easily have been 3's, others could easily have been 1's. This over-compensated for the error by increasing reliability, but in doing so yielded item parameter estimates with mean = .446, standard deviation = 1.40! This suggests that the upper bound of the effect of scoring differences is a mean shift of about .47 from the original mean of .920, with the lower bound being on the order of .095, noted above. (The large difference in standard deviation also illustrates the importance of having comparable levels of reliability across years of testing.)

Lastly, a third set of probabilistically rescored responses was generated using the approach of assigning higher scores conditionally to 85% of the students, and randomly to the remaining 15%. This hybrid approach, modeled to approximate roughly a process whereby most cases where most score increases are genuinely reliable reflections of drifting scorer standards but some are not, yielded item parameter estimates with mean = .752, standard deviation = 1.942. This suggests that an approximation of the effect of scoring differences is a mean shift of about .168 from the original mean of .920. This result is smaller than the additive adjustment obtained by comparing the mean log-odds ratio of the original scores and the rescores, .29, which is in turn well within the upper bound estimate of the effect of scoring differences, .47. (Note that a single outlier, determining by a numerical criterion, was trimmed from each end of the distribution of the differences between log-odds item difficulties in obtaining the log-odds based result for social studies.)

IMPLEMENTING AN APPROPRIATE ADJUSTMENT AT GRADES 4 AND 8

For grades 4 and 8, an inadequate sample of 1994-95 papers was rescored during operational scoring of the 1995-96 papers. Therefore, a special rescoring study was performed on a short timeline. In this study, rescoring was performed on both 1994-95 and 1995-96 papers, using the same set of scorers, a subset of the 1995-96 scorers. Differences between 1995-96 original scores and rescores were compared with differences between 1994-95 original scores and rescores to determine the need for and magnitude of an adjustment to reflect changes in the application of scoring standards.

Due to the need to pull papers efficiently for the special rescoring study, students were sampled rather than papers per se. Instead of attempting to draw a balanced sample of papers with regard to score points for each item in a matrix sampled design, students with the highest total scores (across all items, commmon and matrix-sampled in reading, mathematics, science, social studies, arts & humanities, and practical living/vocational studies) were over-sampled, and a random sample was drawn of remaining students. Papers from the sampled students were scored a second time, and the difference between the recast means on the second scoring (or rescore) and the original score was taken. Positive values for the initial difference statistic, presented in the third column of Table 1 and Table 2, therefore represented relatively lenient rescoring of papers from 1994-95 (hereafter, 1995) or 1995-96 (hereafter, 1996) by the special study scorers. Initial mean differences were compared across the two years (see the final column of Tables 1 and 2) as an indication of instances for which scoring had shifted so dramatically as to merit corrective action.

Recast means for the grade 4 and grade 8 difference scores were computed using the sampling proportions. The weight assigned to each student in each stratum of the sample was equal to the proportion of the effective population represented by the sample divided by the number of students in the stratum of the sample. For rescored 1995-96 common items, weighting was further adjusted to reflect the percentage of papers scored in the established scoring center in Dover, New Hampshire, versus the newly opened scoring center in Methuen, Massachusetts. That is, a weighted mean of the difference score (rescore minus original score) would yield an estimate of the difference that would have been obtained if all papers had been rescored. In the case where one-third of the papers were scored in Dover and two-thirds were scored in Methuen, scores from Dover papers were downweighted by .5, so that the ratio maintained in the sample would be 2 to 1. Tables 1 and 2 present the counts of students in each sample (n) and the number of students they represented (N).

The highest scoring students had been sampled with certainty to represent a relatively small percentage of the effective population, so they received relatively low weights in computing recast means. Because they were sampled with certainty, these students added no error to the mean difference score estimate, which was computed as a composite based on the listed weights in tables 1 and 2 for grades 4 and 8, respectively. For grade 4 reading, for example, the random sample of 1200 students whose 1994-95 papers were rescored had been selected to represent 86.1% of the effective population (students with no scores other than 1, 2, 3 or 4 on any item). Thus, the mean difference score estimate for this population would be .861 times the mean of the difference scores for the random sample, plus .139 times the mean of the difference scores for the certainty sample.

The variance of the mean estimate is the variance of a composite score, i.e., the sum of the variance of the mean estimates multiplied by the square of their weights (a formula which can be found in statistics texts such as Cochran's). For the certainty sample, there is no variance due to sampling error in estimating the mean difference for this stratum, as the entire population of students has had its papers rescored. So, all error in the scores is due to the random sample, and the standard error of the mean difference score is equal to the square root of the following: weight (squared) times the variance of randomly sampled student scores, divided by the number of students in the random sample.

Random sample weights in tables 1 and 2 vary by grade and by year to the extent that the number of students varies across grades and years, and to the extent that a different effective population was used for grade 4, 1995. In this case, students who would have qualified for the IRT calibration sample were included. For the remaining grade and year combinations, students whose papers were to be rescored were drawn only from those who had no blank or zero scores on any item, which is similar to the effective population used in previous years' recasting studies. This is pertinent to the weights used, as a larger effective population serves to weight down the percent of the population sampled with certainty. Sampling also occured by form and, for 1996 items, was limited to four forms per grade. For grade 8, 1995, for example, the effective population was 5920 students.

Once standard errors are calculated for the composite difference scores for each content area, grade and year combination, they are compared with the across year differences of mean difference scores, within content area and grade. If the difference of the mean difference scores exceeds the sum of the two standard errors, then there is no overlap between a one standard error confidence interval around each mean difference score. To elaborate, positive mean difference scores for 1995 indicate the extent to which rescorers are more lenient that original 1995 scorers. Likewise, positive mean difference scores for 1996 indicate the extent to which rescorers are more lenient that original 1996 scorers. If the mean difference score for 1996 exceeds the mean difference score for 1995, then, assuming that the rescorers were consistent in their rescoring of 1995 and 1996 papers, the 1996 original scorers were more stringent than 1995 original scorers. In the case of grade 4 reading, this difference (between scores of -.033 for 1995 and -.017 for 1996) was .016. If -.033 and -.017 were plotted with standard error bands of .012 and .039, respectively, the error bands would overlap. Using a rule-of-thumb noted as having been applied to a profile of differences (Feldt & Brennan, 1989, p. 118), the fact that standard error bands overlap suggests that the two do not differ significantly. It is mathematically equivalent to sum the two standard errors, and to check whether their sum is greater than the absolute value of the difference between the two mean difference scores.

In cases where the difference between the mean difference scores exceeded the sum of the standard error estimates, the following corrective action was taken. Weighted frequencies of item original scores and rescores at the score points 1 through 4 were computed for the sample using the weights noted above. These frequencies were also used to compute log-odds difficulties. After each frequency distribution had been trimmed using a statistical criterion, difficulties based on rescores were compared with difficulties based on original scores. The additive transformation necessary to set the means of the two sets of item difficulties on the same scale was then determined. Transformations based on the 1995 and 1996 log-odds difficulties were then applied to 1995 and 1996 item parameter estimates prior to equating. Scorer adjustments implemented on this basis are presented with the chapter this appendix supplements, in Table 9-3.

CAVEATS

The following notes should be considered. First, standard error estimates do not attempt to model scorer error. That is, the difference statistic, rescore minus score is based on fallible rescores and scores of each paper rather than the 'true' rescore and score. This effect cannot be modeled practicably with the present data. Second, recast mean differences are used as a trigger for applying a correction which is not based explicitly on the recast mean. Instead, a log-odds based approach, capable of detecting differential scoring changes expressed in terms of an additive transformation in a metric that approximates that of item difficulty parameters is employed. Thus, it should not be expected that the magnitude of the recast mean difference will correspond exactly to the magnitude of the impact of applying the corrective log-odds-based transformation. Two alternatives to this bifurcation would be to establish a criterion based on log-odds difficulties (which is not a trivial undertaking), or to use mean differences above the criterion level as both the trigger and as an additive correction (an unacceptable correction, given an insurmountable discrepancy between the raw score and item responses theory-based difficulty estimates). The third alternative, applying corrections in all cases, will be given further consideration for future use.

Table 1
Grade 4 Rescore Weighted Mean Difference Statistics

Content Area	Year, Sample	Proportion of Papers Scored At Each Site (If Applicable)	Statistics Describing Difference Between Original Scores and Rescores of Papers from Each Year for Overall Sample and Stratum Sampled Without Certainty					Composite Contribution			Net Year-To-Year Difference of Rescore Means
			Mean	Std. Dev.	Variance	n	N Represented	Weight	Error Var.	Std. Error	
Reading	1995(Total)		-.033	.505	.255		10629			.012	
	1995(Sampled)		-.027	.501	.251	1200	9147	.861	.00015		.016
	1996(Total)		-.017	.403	.162					.039	
	1996(Sampled)		-.017	.400	.160	100		.986	.00155		
Math	1995(Total)		-.025	.440	.193		10629			.011	
	1995(Sampled)		-.022	.445	.198	1200	9147	.861	.00012		-.005
	1996(Total)		-.030	.410	.168					.040	
	1996(Sampled)		-.031	.410	.168	100		.986	.00163		
Science	1995(Total)		-.049	.445	.198		10629			.011	
	1995(Sampled)		-.048	.447	.200	1200	9147	.861	.00012		.058
	1996(Total)		.009	.501	.251				.00117	.034	
	Methuen Total	.333	-.039	.633	.401		7928				
	Dover Total	.667	.034	.415	.172		7871				
	Methuen Sample		-.036	.634	.401	100	7801	.328	.00043		
	Dover Sample		.036	.413	.171	100	7758	.657	.00074		
Soc. St.	1995(Total)		-.067	.534	.285		10629			.013	
	1995(Sampled)		-.062	.528	.279	1200	9147	.861	.00017		.071
	1996(Total)		.004	.620	.385				.00183	.043	
	Methuen Total	.333	-.116	.748	.559		7928				
	Dover Total	.667	.063	.536	.287		7871				
	Methuen Sample		-.109	.746	.557	100	7801	.328	.00060		
	Dover Sample		.067	.534	.285	100	7758	.657	.00123		
Arts/Hum	1995(Total)		.048	.530	.281		10629			.013	
	1995(Sampled)	.667	.055	.520	.271	1200	9147	.861	.00017		-.057
	1996(Estimated)	.333	-.009	.491	.241				.00150	.039	
PLVS	1995(Total)		-.032	.561	.315		10629			.014	
	1995(Sampled)		-.028	.555	.308	1200	9147	.861	.00019		-.023
	1996(Estimated)		-.009	.491	.241				.00150	.039	
Note: see text for further description of column information and calculations											

Table 2
Grade 8 Rescore Weighted Mean Difference Statistics

Content Area	Year, Sample	Proportion of Papers Scored At Each Site (If Applicable	Statistics Describing Difference Between Original Scores and Rescores of Papers from Each Year for Overall Sample and Stratum Sampled Without Certainty					Composite Contribution			Net Year-To-Year Difference of Rescor Means
			Mean	Std. Dev.	Variance	n	N Represented	Weight	Error Var.	Std. Error	
Reading	1995(Total)		.016	.456	.208		5920			.011	
	1995(Sampled)		.015	.462	.214	1200	4873	.823	.00012		-.025
	1996(Total)		-.009	.464	.215				.00083	.029	
	Methuen Total	.667	.006	.463	.214		1355				
	Dover Total	.333	-.041	.464	.215		778				
	Methuen Sample		.007	.451	.204	100	1183	.582	.00069		
	Dover Sample		-.027	.456	.208	100	608	.260	.00014		
Math	1995(Total)		-.017	.414	.172		5920			.010	
	1995(Sampled)		-.011	.411	.169	1200	4873	.823	.00010		.017
	1996(Total)		-.001	.489	.239				.00116	.034	
	Methuen Total	.750	.000	.502	.252		1355				
	Dover Total	.250	-.005	.447	.200		687				
	Methuen Sample		.003	.507	.257	100	1183	.655	.00110		
	Dover Sample		.013	.445	.198	105	486	.177	.00006		
Science	1995(Total)		-.094	.502	.252		5920			.012	
	1995(Sampled)		-.080	.492	.242	1200	4873	.823	.00014		-.011
	1996(Total)		-.106	.588	.346				.00164	.041	
	Methuen Total	.750	-.124	.612	.375		1355				
	Dover Total	.250	-.051	.504	.254		687				
	Methuen Sample		-.119	.605	.366	100	1183	.655	.00157		
	Dover Sample		-.064	.494	.244	105	486	.177	.00007		
Soc. St.	1995(Total)		-.108	.621	.386		5920			.015	
	1995(Sampled)		-.089	.613	.376	1200	4873	.823	.00021		.074
	1996(Total)		-.034	.560	.314				.00126	.035	
	Methuen Total	.667	-.005	.571	.326		1355				
	Dover Total	.333	-.093	.533	.284		778				
	Methuen Sample		-.002	.565	.319	100	1183	.582	.00108		
	Dover Sample		-.069	.509	.259	100	608	.260	.00018		
Arts/Hum	1995(Total)		-.150	.572	.327		5920			.013	
	1995(Sampled)		-.142	.561	.314	1200	4873	.823	.00018		.102
	1996(Estimated)		-.047	.488	.239				.00122	.035	
PLVS	1995(Total)		.023	.475	.226		5920			.011	
	1995(Sampled)		.029	.464	.216	1200	4873	.823	.00012		-.070
	1996(Estimated)		-.047	.488	.239				.00122	.035	

Note: see text for further description of column information and calculations

Appendix 9-3
Locations of Open-Response Linking Items by Year

LINKING ITEMS FOR GRADE 4 READING

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 1 / Item 6	Form 1 / Item 6	
Form 1 / Item 7	Form 5 / Item 7	
Form 3 / Item 7	Form 3 / Item 6	Form 3 / Item 7
Form 4 / Item 7	Form 4 / Item 7	
Form 6 / Item 6	Form 6 / Item 6	Form 10 / Item 7
Form 7 / Item 7	Form 7 / Item 7	Form 7 / Item 6
Form 8 / Item 6	Form 8 / Item 6	
Form 9 / Item 7	Form 8 / Item 7	Form 4 / Item 6
Form 10 / Item 6	Form 9 / Item 6	Form 2 / Item 7
Form 10 / Item 7	Form 10 / Item 7	Form 7 / Item 7
Form 12 / Item 7	Form 12 / Item 7	Form 5 / Item 6
	Common / Item 1	Form 11 / Item 6
	Common / Item 2	Form 11 / Item 7
	Common / Item 3	Form 12 / Item 7
	Common / Item 4	Form 9 / Item 7
	Common / Item 5	Form 10 / Item 6
	Form 2 / Item 6	Form 1 / Item 7
	Form 2 / Item 6	Form 1 / Item 6
	Form 3 / Item 7	Form 3 / Item 6
	Form 12 / Item 6	Form 5 / Item 7

LINKING ITEMS FOR GRADE 4 MATHEMATICS

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 3 / Item 7	Form 11 / Item 7	
Form 5 / Item 7	Form 4 / Item 7	
Form 6 / Item 6	Form 6 / Item 7	Form 6 / Item 7
Form 6 / Item 7	Form 6 / Item 6	
Form 7 / Item 6	Form 12 / Item 7	Form 5 / Item 7
Form 7 / Item 7	Form 7 / Item 7	Form 12 / Item 6
Form 8 / Item 7	Form 7 / Item 6	
Form 9 / Item 6	Form 9 / Item 6	Form 9 / Item 6
Form 9 / Item 7	Form 9 / Item 7	Form 5 / Item 6
Form 10 / Item 7	Form 1 / Item 6	
Form 11 / Item 6	Form 12 / Item 6	
Form 12 / Item 7	Form 10 / Item 7	
	Common / Item 1	Form 2 / Item 7
	Common / Item 2	Form 6 / Item 6
	Common / Item 3	Form 10 / Item 7
	Common / Item 4	Form 4 / Item 7
	Form 1 / Item 7	Form 7 / Item 6
	Form 2 / Item 7	Form 1 / Item 6
	Form 2 / Item 7	Form 1 / Item 7
	Form 5 / Item 7	Form 9 / Item 7
	Form 10 / Item 6	Form 3 / Item 7
	Form 11 / Item 6	Form 12 / Item 7

LINKING ITEMS FOR GRADE 4 SCIENCE

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 2 / Item 6	Form 3 / Item 6	Form 1 / Item 6
Form 6 / Item 6	Form 11 / Item 6	Form 10 / Item 6
Form 7 / Item 6	Form 1 / Item 7	Form 1 / Item 7
Form 7 / Item 7	Form 7 / Item 7	
Form 8 / Item 6	Form 10 / Item 6	
Form 9 / Item 6	Form 4 / Item 7	
Form 9 / Item 7	Form 4 / Item 6	Form 8 / Item 6
Form 10 / Item 6	Form 5 / Item 6	
Form 11 / Item 6	Form 9 / Item 6	Form 5 / Item 6
Form 11 / Item 7	Form 9 / Item 7	
Form 12 / Item 6	Form 8 / Item 6	
Form 12 / Item 7	Form 8 / Item 7	Form 11 / Item 7
	Common / Item 1	Form 3 / Item 7
	Common / Item 2	Form 9 / Item 6
	Common / Item 3	Form 9 / Item 7
	Common / Item 4	Form 12 / Item 6
	Common / Item 5	Form 8 / Item 7
	Form 2 / Item 6	Form 7 / Item 6
	Form 5 / Item 7	Form 10 / Item 7
	Form 6 / Item 6	Form 11 / Item 6
	Form 10 / Item 7	Form 4 / Item 7

LINKING ITEMS FOR GRADE 4 SOCIAL STUDIES

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 1 / Item 7	Form 1 / Item 7	
Form 3 / Item 6	Form 12 / Item 6	
Form 3 / Item 7	Form 6 / Item 7	Form 5 / Item 6
Form 6 / Item 6	Form 6 / Item 6	
Form 7 / Item 7	Form 11 / Item 6	
Form 9 / Item 7	Form 11 / Item 7	Form 4 / Item 7
Form 11 / Item 6	Form 10 / Item 6	
Form 11 / Item 7	Form 10 / Item 7	
	Common / Item 1	Form 7 / Item 6
	Common / Item 4	Form 8 / Item 6
	Common / Item 5	Form 2 / Item 7
	Form 1 / Item 6	Form 1 / Item 6
	Form 3 / Item 6	Form 3 / Item 7
	Form 3 / Item 7	Form 3 / Item 6
	Form 4 / Item 6	Form 11 / Item 6
	Form 5 / Item 7	Form 7 / Item 7
	Form 8 / Item 6	Form 10 / Item 6
	Form 9 / Item 7	Form 9 / Item 7
	Form 12 / Item 7	Form 6 / Item 7

LINKING ITEMS FOR GRADE 4 ARTS & HUMANITIES			
1995-96 LOCATION	1994-95 LOCATION*	1993-94 LOCATION**	1992-93 LOCATION**
Form 2 / Item 1	Form 12 / Item 2		
Form 4 / Item 1	Form 4 / Item 1	Form 4 / Item 7 (R)	Form 10 / Item 7 (R)
Form 6 / Item 2	Form 1 / Item 1		
Form 6 / Item 2	Form 12 / Item 1		
Form 9 / Item 2	Form 4 / Item 2	Form 10 / Item 7 (H)	Form 10 / Item 7 (H)
Form 10 / Item 1	Form 5 / Item 2	Form 11 / Item 7 (H)	Form 11 / Item 7 (H)
Form 10 / Item 1	Form 11 / Item 2	Form 11 / Item 7 (H)	Form 11 / Item 7 (H)
Form 12 / Item 1	Form 2 / Item 1		
Form 12 / Item 1	Form 8 / Item 1		
	Form 10 / Item 1	Form 4 / Item 7 (R)	Form 10 / Item 7 (R)
		Form 5 / Item 7 (H)	Form 8 / Item 7 (H)
		Form 6 / Item 7 (S)	Form 8 / Item 7 (S)
	Form 7 / Item 1	Form 7 / Item 7 (M)	Form 12 / Item 7 (M)
	Form 9 / Item 1	Form 7 / Item 7 (M)	Form 12 / Item 7 (M)
	Form 10 / Item 2	Form 10 / Item 7 (H)	Form 10 / Item 7 (H)
		Form 12 / Item 6 (R)	Form 9 / Item 6 (R)

* In 1994-95, the following forms contained the same pairing of AH and PL items: forms 1 and 12, forms 2 and 8, forms 3 and 6, forms 4 and 10, forms 5 and 11, and forms 7 and 9.

** In 1992-93 and 1993-94, Arts and Humanities and Practical Living/Vocational Studies items were administered as part of the open-response Reading, Mathematics, Science and Social Studies tests. The test associated with an item is denoted by one of the following:

(R) Reading
(M) Mathematics
(S) Science
(H) Social Studies.

LINKING ITEMS FOR GRADE 4 PRACTICAL LIVING/VOCATIONAL STUDIES

1995-96 LOCATION	1994-95 LOCATION*	1993-94 LOCATION**	1992-93 LOCATION**
Form 2 / Item 4	Form 2 / Item 3		
Form 2 / Item 4	Form 8 / Item 3		
Form 3 / Item 3	Form 7 / Item 4		
Form 3 / Item 3	Form 9 / Item 4		
Form 5 / Item 4	Form 7 / Item 3	Form 8 / Item 7 (M)	Form 6 / Item 7 (M)
Form 5 / Item 4	Form 9 / Item 3	Form 8 / Item 7 (M)	Form 6 / Item 7 (M)
Form 6 / Item 4	Form 4 / Item 4		
Form 6 / Item 4	Form 10 / Item 4		
Form 7 / Item 3	Form 2 / Item 4		
Form 7 / Item 3	Form 8 / Item 4		
	Form 4 / Item 3	Form 1 / Item 7 (H)	Form 9 / Item 7 (H)
	Form 10 / Item 3	Form 1 / Item 7 (H)	Form 9 / Item 7 (H)
		Form 2 / Item 6 (M)	Form 2 / Item 6 (M)
	Form 1 / Item 3	Form 4 / Item 6 (S)	Form 9 / Item 6 (S)
	Form 12 / Item 3	Form 4 / Item 6 (S)	Form 9 / Item 6 (S)
	Form 5 / Item 3	Form 5 / Item 7 (S)	Form 5 / Item 7 (S)
	Form 11 / Item 3	Form 5 / Item 7 (S)	Form 5 / Item 7 (S)
		Form 6 / Item 6 (R)	Form 12 / Item 6 (R)
	Form 1 / Item 4	Form 9 / Item 6 (R)	Form 2 / Item 7 (R)
	Form 12 / Item 4	Form 9 / Item 6 (R)	Form 2 / Item 7 (R)
	Form 3 / Item 3	Form 10 / Item 6 (M)	
	Form 6 / Item 3	Form 10 / Item 6 (M)	
		Form 11 / Item 6 (M)	Form 11 / Item 6 (M)
		Form 12 / Item 7 (H)	Form 12 / Item 7 (H)

* In 1994-95, the following forms contained the same pairing of AH and PL items: forms 1 and 12, forms 2 and 8, forms 3 and 6, forms 4 and 10, forms 5 and 11, and forms 7 and 9.

** In 1992-93 and 1993-94, Arts and Humanities and Practical Living/Vocational Studies items were administered as part of the open-response Reading, Mathematics, Science and Social Studies tests. The test associated with an item is denoted by one of the following:
(R) Reading
(M) Mathematics
(S) Science
(H) Social Studies.

LINKING ITEMS FOR GRADE 8 READING

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 1 / Item 8	Form 1 / Item 8	
Form 3 / Item 7	Form 12 / Item 8	Form 12 / Item 7
Form 4 / Item 7	Form 10 / Item 7	
Form 4 / Item 8	Form 10 / Item 8	
Form 5 / Item 7	Form 5 / Item 8	
Form 5 / Item 8	Form 2 / Item 7	
Form 6 / Item 7	Form 9 / Item 8	Form 10 / Item 7
Form 7 / Item 8	Form 7 / Item 7	
Form 8 / Item 7	Form 7 / Item 8	Form 5 / Item 6
Form 10 / Item 8	Form 6 / Item 7	Form 11 / Item 7
Form 12 / Item 7	Form 6 / Item 8	
	Common / Item 1	Form 10 / Item 6
	Common / Item 3	Form 8 / Item 7
	Common / Item 4	Form 1 / Item 6
	Common / Item 5	Form 8 / Item 6
	Common / Item 6	Form 3 / Item 6
	Form 3 / Item 7	Form 3 / Item 7
	Form 4 / Item 7	Form 4 / Item 6
	Form 8 / Item 8	Form 2 / Item 6
	Form 9 / Item 7	Form 7 / Item 6
	Form 11 / Item 7	Form 11 / Item 6
	Form 12 / Item 7	Form 6 / Item 6

LINKING ITEMS FOR GRADE 8 MATHEMATICS

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 2 / Item 7	Form 2 / Item 7	Form 10 / Item 7
Form 3 / Item 8	Form 3 / Item 8	
Form 4 / Item 7	Form 4 / Item 7	Form 9 / Item 6
Form 4 / Item 8	Form 4 / Item 8	
Form 5 / Item 7	Form 5 / Item 7	Form 7 / Item 6
Form 5 / Item 8	Form 7 / Item 8	
Form 6 / Item 8	Form 6 / Item 8	Form 2 / Item 7
Form 7 / Item 7	Form 7 / Item 7	Form 11 / Item 6
Form 8 / Item 7	Form 8 / Item 7	
Form 9 / Item 7	Form 9 / Item 8	Form 9 / Item 7
Form 10 / Item 7	Form 10 / Item 7	
Form 11 / Item 8	Form 12 / Item 7	Form 3 / Item 6
	Common / Item 1	Form 10 / Item 6
	Common / Item 2	Form 6 / Item 6
	Common / Item 3	Form 12 / Item 7
	Common / Item 5	Form 4 / Item 7
	Common / Item 6	Form 6 / Item 7
	Form 6 / Item 7	Form 5 / Item 7
	Form 11 / Item 7	Form 1 / Item 6
	Form 11 / Item 8	Form 5 / Item 6

LINKING ITEMS FOR GRADE 8 SCIENCE

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 1 / Item 7	Form 1 / Item 7	
Form 2 / Item 7	Form 2 / Item 7	Form 9 / Item 6
Form 3 / Item 7	Form 3 / Item 7	
Form 8 / Item 7	Form 8 / Item 7	
Form 9 / Item 8	Form 11 / Item 7	Form 8 / Item 6
Form 10 / Item 7	Form 10 / Item 7	
Form 11 / Item 8	Form 7 / Item 8	Form 7 / Item 6
Form 12 / Item 8	Form 5 / Item 8	Form 3 / Item 7
	Common / Item 1	Form 3 / Item 6
	Common / Item 2	Form 11 / Item 7
	Common / Item 3	Form 1 / Item 6
	Common / Item 4	Form 8 / Item 7
	Common / Item 5	Form 9 / Item 7
	Common / Item 6	Form 6 / Item 6
	Form 1 / Item 8	Form 1 / Item 7
	Form 4 / Item 8	Form 4 / Item 7
	Form 4 / Item 8	Form 4 / Item 7
	Form 7 / Item 7	Form 7 / Item 7
	Form 9 / Item 8	Form 12 / Item 6
	Form 10 / Item 8	Form 2 / Item 6
	Form 11 / Item 8	Form 5 / Item 7
	Form 12 / Item 8	Form 10 / Item 7

LINKING ITEMS FOR GRADE 8 SOCIAL STUDIES		
1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 1 / Item 7	Form 1 / Item 7	
Form 1 / Item 8	Form 1 / Item 8	
Form 2 / Item 7	Form 5 / Item 7	
Form 3 / Item 7	Form 3 / Item 7	Form 4 / Item 6
Form 5 / Item 7	Form 9 / Item 7	Form 9 / Item 6
Form 6 / Item 7	Form 4 / Item 7	
Form 7 / Item 7	Form 11 / Item 7	
Form 7 / Item 8	Form 11 / Item 8	Form 7 / Item 7
Form 8 / Item 7	Form 6 / Item 7	
Form 8 / Item 8	Form 6 / Item 8	
Form 11 / Item 7	Form 2 / Item 8	Form 1 / Item 7
	Common / Item 1	Form 7 / Item 6
	Common / Item 2	Form 5 / Item 6
	Common / Item 3	Form 11 / Item 6
	Common / Item 4	Form 8 / Item 6
	Common / Item 5	Form 9 / Item 7
	Form 3 / Item 8	Form 4 / Item 7
	Form 4 / Item 8	Form 10 / Item 7
	Form 5 / Item 8	Form 11 / Item 7
	Form 7 / Item 8	Form 2 / Item 7
	Form 10 / Item 7	Form 12 / Item 6

LINKING ITEMS FOR GRADE 8 ARTS & HUMANITIES

1995-96 LOCATION	1994-95 LOCATION*	1993-94 LOCATION**	1992-93 LOCATION**
Form 4 / Item 1	Form 1 / Item 1		
Form 4 / Item 1	Form 2 / Item 1		
Form 4 / Item 2	Form 4 / Item 2	Form 11 / Item 6 (S)	Form 10 / Item 6 (S)
Form 4 / Item 2	Form 5 / Item 2	Form 11 / Item 6 (S)	Form 10 / Item 6 (S)
Form 6 / Item 1	Form 6 / Item 2	Form 12 / Item 7 (H)	Form 3 / Item 6 (H)
Form 6 / Item 1	Form 11 / Item 2	Form 12 / Item 7 (H)	Form 3 / Item 6 (H)
Form 9 / Item 1	Form 1 / Item 2	Form 2 / Item 7 (S)	Form 4 / Item 7 (S)
Form 9 / Item 1	Form 2 / Item 2	Form 2 / Item 7 (S)	Form 4 / Item 7 (S)
Form 11 / Item 1	Form 3 / Item 2	Form 10 / Item 6 (H)	Form 6 / Item 6 (H)
Form 11 / Item 1	Form 8 / Item 2	Form 10 / Item 6 (H)	Form 6 / Item 6 (H)
Form 12 / Item 1	Form 9 / Item 1	Form 1 / Item 7 (R)	
Form 12 / Item 1	Form 12 / Item 1	Form 1 / Item 7 (R)	
	Form 3 / Item 1	Form 4 / Item 6 (M)	Form 1 / Item 6 (M)
	Form 8 / Item 1	Form 4 / Item 6 (M)	Form 1 / Item 6 (M)
		Form 8 / Item 7 (H)	Form 7 / Item 7 (H)

* In 1994-95, the following forms contained the same pairing of AH and PL items: forms 1 and 2, forms 3 and 8, forms 4 and 5, forms 6 and 11, forms 7 and 10, and forms 9 and 12.

** In 1992-93 and 1993-94, Arts and Humanities and Practical Living/Vocational Studies items were administered as part of the open-response Reading, Mathematics, Science and Social Studies tests. The test associated with an item is denoted by one of the following:

(R) Reading
(M) Mathematics
(S) Science
(H) Social Studies.

LINKING ITEMS FOR GRADE 8 PRACTICAL LIVING/VOCATIONAL STUDIES			
1995-96 LOCATION	1994-95 LOCATION*	1993-94 LOCATION**	1992-93 LOCATION**
Form 1 / Item 4	Form 9 / Item 4		
Form 1 / Item 4	Form 12 / Item 4		
Form 4 / Item 4	Form 1 / Item 4		
Form 4 / Item 4	Form 2 / Item 4		
Form 6 / Item 4	Form 6 / Item 3	Form 5 / Item 7 (R)	Form 5 / Item 7 (R)
Form 6 / Item 4	Form 11 / Item 3	Form 5 / Item 7 (R)	Form 5 / Item 7 (R)
Form 8 / Item 3	Form 3 / Item 3		
Form 8 / Item 3	Form 8 / Item 3		
Form 10 / Item 3	Form 6 / Item 4		
Form 10 / Item 3	Form 11 / Item 4		
Form 12 / Item 3	Form 7 / Item 4		
Form 12 / Item 3	Form 10 / Item 4		
		Form 1 / Item 6 (H)	Form 12 / Item 6 (H)
	Form 7 / Item 3	Form 2 / Item 6 (M)	Form 10 / Item 6 (M)
	Form 10 / Item 3	Form 2 / Item 6 (M)	Form 10 / Item 6 (M)
	Form 3 / Item 4	Form 5 / Item 7 (H)	Form 6 / Item 7 (H)
	Form 8 / Item 4	Form 5 / Item 7 (H)	Form 5 / Item 7 (H)
	Form 1 / Item 3	Form 6 / Item 7 (S)	Form 11 / Item 7 (S)
	Form 2 / Item 3	Form 6 / Item 7 (S)	Form 11 / Item 7 (S)
	Form 4 / Item 4	Form 7 / Item 7 (M)	Form 7 / Item 7 (M)
	Form 5 / Item 4	Form 7 / Item 7 (M)	Form 7 / Item 7 (M)
		Form 8 / Item 7 (M)	Form 8 / Item 7 (M)
		Form 10 / Item 6 (S)	Form 9 / Item 6 (S)
	Form 9 / Item 3	Form 11 / Item 7 (M)	Form 6 / Item 7 (M)
	Form 12 / Item 3	Form 11 / Item 7 (M)	Form 6 / Item 7 (M)

* In 1994-95, the following forms contained the same pairing of AH and PL items: forms 1 and 2, forms 3 and 8, forms 4 and 5, forms 6 and 11, forms 7 and 10, and forms 9 and 12.

** In 1992-93 and 1993-94, Arts and Humanities and Practical Living/Vocational Studies items were administered as part of the open-response Reading, Mathematics, Science and Social Studies tests. The test associated with an item is denoted by one of the following:

(R) Reading
(M) Mathematics
(S) Science
(H) Social Studies.

LINKING ITEMS FOR GRADE 11/12 READING

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Common / Item 6	Form 10 / Item 8	Form 1 / Item 6
Form 1 / Item 8	Form 2 / Item 7	Form 2 / Item 6
Form 2 / Item 8	Form 1 / Item 7	
Form 7 / Item 7	Form 5 / Item 7	
Form 7 / Item 8	Form 6 / Item 8	
Form 9 / Item 7	Form 9 / Item 7	
Form 9 / Item 8	Form 3 / Item 7	
Form 11 / Item 7	Form 10 / Item 8	Form 10 / Item 7
Form 11 / Item 8	Form 11 / Item 8	
Form 12 / Item 7	Form 12 / Item 7	Form 12 / Item 6
Form 12 / Item 8	Form 7 / Item 8	
	Common / Item 1	Form 5 / Item 7
	Common / Item 2	Form 9 / Item 6
	Common / Item 3	Form 8 / Item 6
	Common / Item 4	Form 11 / Item 7
	Common / Item 5	Form 4 / Item 6
	Common / Item 6	Form 10 / Item 6
	Form 2 / Item 8	Form 2 / Item 7
	Form 3 / Item 8	Form 12 / Item 7
	Form 4 / Item 8	Form 1 / Item 7
	Form 6 / Item 7	Form 6 / Item 6
	Form 7 / Item 7	Form 7 / Item 6

LINKING ITEMS FOR GRADE 11/12 MATHEMATICS

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 2 / Item 7	Form 2 / Item 7	Form 12 / Item 6
Form 3 / Item 8	Form 3 / Item 8	Form 1 / Item 7
Form 4 / Item 7	Form 4 / Item 7	Form 4 / Item 6
Form 4 / Item 8	Form 4 / Item 8	
Form 6 / Item 7	Form 6 / Item 7	Form 6 / Item 6
Form 8 / Item 8	Form 8 / Item 8	Form 8 / Item 7
Form 10 / Item 7	Form 10 / Item 7	Form 10 / Item 7
Form 10 / Item 8	Form 10 / Item 8	
Form 12 / Item 7	Form 9 / Item 7	Form 9 / Item 6
	Common / Item 1	Form 10 / Item 6
	Common / Item 2	Form 6 / Item 7
	Common / Item 3	Form 2 / Item 6
	Common / Item 4	Form 11 / Item 7
	Common / Item 6	Form 9 / Item 7
	Form 1 / Item 7	Form 11 / Item 6
	Form 5 / Item 7	Form 5 / Item 6
	Form 6 / Item 8	Form 3 / Item 6
	Form 8 / Item 7	Form 8 / Item 6

LINKING ITEMS FOR GRADE 11/12 SCIENCE

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 3 / Item 7	Form 3 / Item 7	
Form 3 / Item 8	Form 3 / Item 8	
Form 6 / Item 8	Form 4 / Item 7	
Form 7 / Item 7	Form 7 / Item 8	Form 8 / Item 7
Form 9 / Item 7	Form 9 / Item 7	
Form 9 / Item 8	Form 4 / Item 8	
Form 10 / Item 7	Form 10 / Item 7	Form 10 / Item 6
Form 11 / Item 7	Form 11 / Item 7	
Form 11 / Item 8	Form 11 / Item 8	
	Common / Item 1	Form 11 / Item 6
	Common / Item 3	Form 7 / Item 6
	Common / Item 4	Form 4 / Item 6
	Common / Item 5	Form 12 / Item 7
	Common / Item 6	Form 1 / Item 7
	Form 1 / Item 7	Form 2 / Item 6
	Form 2 / Item 7	Form 1 / Item 6
	Form 2 / Item 8	Form 2 / Item 7
	Form 5 / Item 7	Form 5 / Item 6
	Form 5 / Item 8	Form 5 / Item 7
	Form 6 / Item 8	Form 6 / Item 7
	Form 6 / Item 8	Form 6 / Item 7
	Form 7 / Item 7	Form 7 / Item 7
	Form 10 / Item 8	Form 11 / Item 7
	Form 12 / Item 7	Form 12 / Item 6

LINKING ITEMS FOR GRADE 11/12 SOCIAL STUDIES

1995-96 LOCATION	1994-95 LOCATION	1993-94 LOCATION
Form 2 / Item 7	Form 5 / Item 8	
Form 3 / Item 7	Form 3 / Item 7	Form 12 / Item 6
Form 3 / Item 8	Form 3 / Item 8	
Form 4 / Item 7	Form 4 / Item 7	
Form 5 / Item 7	Form 1 / Item 7	Form 1 / Item 6
Form 5 / Item 8	Form 4 / Item 8	
Form 8 / Item 7	Form 2 / Item 8	Form 6 / Item 6
Form 9 / Item 7	Form 9 / Item 7	Form 11 / Item 6
Form 9 / Item 8	Form 9 / Item 8	Form 9 / Item 7
Form 10 / Item 7	Form 10 / Item 7	Form 10 / Item 7
Form 10 / Item 7	Form 10 / Item 7	Form 10 / Item 6
Form 11 / Item 8	Form 11 / Item 8	
Form 12 / Item 7	Form 12 / Item 7	
Form 12 / Item 8	Form 12 / Item 8	Form 3 / Item 7
	Common / Item 1	Form 4 / Item 6
	Common / Item 2	Form 8 / Item 7
	Common / Item 3	Form 8 / Item 6
	Common / Item 4	Form 5 / Item 7
	Common / Item 5	Form 2 / Item 7
	Form 1 / Item 8	Form 1 / Item 7
	Form 2 / Item 7	Form 3 / Item 6

LINKING ITEMS FOR GRADE 11/12 ARTS & HUMANITIES

1995-96 LOCATION	1994-95 LOCATION*	1993-94 LOCATION**	1992-93 LOCATION**
Form 4 / Item 1	Form 8 / Item 1	Form 3 / Item 7 (R)	
Form 4 / Item 1	Form 9 / Item 1	Form 3 / Item 7 (R)	
Form 6 / Item 1	Form 4 / Item 2	Form 9 / Item 7 (S)	Form 4 / Item 7 (S)
Form 6 / Item 1	Form 6 / Item 2	Form 9 / Item 7 (S)	Form 4 / Item 7 (S)
Form 7 / Item 1	Form 7 / Item 1		
Form 7 / Item 1	Form 10 / Item 1		
Form 9 / Item 2	Form 3 / Item 2		
Form 9 / Item 2	Form 5 / Item 2		
Form 10 / Item 1	Form 7 / Item 2		
Form 10 / Item 1	Form 10 / Item 2		
Form 11 / Item 2	Form 4 / Item 1	Form 12 / Item 7 (H)	Form 12 / Item 7 (H)
Form 11 / Item 2	Form 6 / Item 1	Form 12 / Item 7 (H)	Form 12 / Item 7 (H)
Form 12 / Item 1	Form 1 / Item 1	Form 8 / Item 7 (R)	Form 8 / Item 7 (R)
Form 12 / Item 1	Form 12 / Item 1	Form 8 / Item 7 (R)	Form 8 / Item 7 (R)
		Form 4 / Item 7 (M)	Form 10 / Item 7 (M)
	Form 1 / Item 2	Form 6 / Item 7 (H)	Form 11 / Item 7 (H)
	Form 12 / Item 2	Form 6 / Item 7 (H)	Form 11 / Item 7 (H)
	Form 3 / Item 1	Form 10 / Item 7 (S)	Form 10 / Item 7 (S)
	Form 5 / Item 1	Form 10 / Item 7 (S)	Form 10 / Item 7 (S)
		Common / Item 7 (R)	Common / Item 7 (R)

* In 1994-95, the following forms contained the same pairing of AH and PL items: forms 1 and 12, forms 2 and 11, forms 3 and 5, forms 4 and 6, forms 7 and 10, and forms 8 and 9.

** In 1992-93 and 1993-94, Arts and Humanities and Practical Living/Vocational Studies items were administered as part of the open-response Reading, Mathematics, Science and Social Studies tests. The test associated with an item is denoted by one of the following:

(R) Reading
(M) Mathematics
(S) Science
(H) Social Studies.

LINKING ITEMS FOR GRADE 11/12 PRACTICAL LIVING/VOCATIONAL STUDIES

1995-96 LOCATION	1994-95 LOCATION*	1993-94 LOCATION**	1992-93 LOCATION**
Form 2 / Item 3	Form 8 / Item 3		
Form 2 / Item 3	Form 9 / Item 3		
Form 4 / Item 3	Form 3 / Item 3		
Form 4 / Item 3	Form 5 / Item 3		
Form 6 / Item 3	Form 4 / Item 4	Form 9 / Item 7 (R)	Form 9 / Item 7 (R)
Form 6 / Item 3	Form 6 / Item 4	Form 9 / Item 7 (R)	Form 9 / Item 7 (R)
Form 7 / Item 4	Form 7 / Item 4	Form 4 / Item 7 (S)	Form 9 / Item 7 (S)
Form 7 / Item 4	Form 10 / Item 4	Form 4 / Item 7 (S)	Form 9 / Item 7 (S)
Form 8 / Item 3	Form 2 / Item 3		
Form 8 / Item 3	Form 11 / Item 3		
Form 9 / Item 3	Form 4 / Item 3		
Form 9 / Item 3	Form 6 / Item 3		
Form 10 / Item 4	Form 2 / Item 4	Form 7 / Item 7 (H)	
Form 10 / Item 4	Form 11 / Item 4	Form 7 / Item 7 (H)	
Form 12 / Item 3	Form 7 / Item 3		
Form 12 / Item 3	Form 10 / Item 3		
		Form 2 / Item 7 (M)	Form 7 / Item 7 (M)
	Form 8 / Item 4	Form 3 / Item 7 (S)	Form 3 / Item 7 (S)
	Form 9 / Item 4	Form 3 / Item 7 (S)	Form 3 / Item 7 (S)
		Form 8 / Item 6 (S)	Form 8 / Item 6 (S)
	Form 3 / Item 4	Form 11 / Item 7 (H)	Form 6 / Item 7 (H)
	Form 5 / Item 4	Form 11 / Item 7 (H)	Form 6 / Item 7 (H)
	Form 1 / Item 4	Form 12 / Item 7 (M)	Form 12 / Item 7 (M)
	Form 12 / Item 4	Form 12 / Item 7 (M)	Form 12 / Item 7 (M)

* In 1994-95, the following forms contained the same pairing of AH and PL items: forms 1 and 12, forms 2 and 11, forms 3 and 5, forms 4 and 6, forms 7 and 10, and forms 8 and 9.

** In 1992-93 and 1993-94, Arts and Humanities and Practical Living/Vocational Studies items were administered as part of the open-response Reading, Mathematics, Science and Social Studies tests. The test associated with an item is denoted by one of the following:

(R) Reading
(M) Mathematics
(S) Science
(H) Social Studies.

Appendix 9-4
Open-Response Performance Level Cutpoints

Performance Level Cut Points

		GRADE	
CONTENT AREA	**4**	**8**	**11/12**
Reading*	-.47326	-.91719	0.02080
	1.25162	1.11481	1.55008
	2.51874	2.44343	2.72838
Mathematics	0.17470	0.17945	0.02237
	1.31129	0.88822	0.98002
	1.63545	1.32255	1.59604
Science	0.16885	0.02015	-.44274
	1.75488	1.75251	1.55279
	2.67713	2.84953	2.43649
Social Studies	-.14454	-.28538	-.09521
	1.18513	1.14255	1.33665
	2.43169	2.33213	2.57729

* Note: For each grade and content area, the three values listed are the cut-points for Novice vs. Apprentice, Apprentice vs. Proficient, and Proficient vs. Distinguished, respectively.

Appendix 9-5
Judgment-Based Performance Events Equating Study

JUDGMENT-BASED EQUATING OF PERFORMANCE EVENTS

Abstract

The purpose of this study was to place results from the 1994-95 KIRIS Performance Events testing on the same scale as the results from the 1991-92, 1992-93, and 1993-94 administrations of the test. A judgment-based procedure was used to link the scales. Judges were asked to determine whether performance at each 1994-95 score points falls above or below the Novice/Apprentice, Apprentice/Proficient, or Proficient/Distinguished performance level cutpoints established for the 1993-94 performance events.

The study was conducted in two phases. In Phase I, examples of student work from 1993-94 were selected to create the cutpoint benchmark folders against which 1994-95 performance would be compared. In Phase II, judges categorized each 1994-95 raw score point as Novice, Apprentice, Proficient, or Distinguished by comparing the level of performance exhibited on a sample of papers at that score point with the level of performance exhibited in the benchmark folders.

Background

Each spring, performance events are administered to Kentucky students in grades 4, 8, and 11[1] as part of the annual KIRIS assessment. Since 1991-92 performance events have been administered in mathematics, science, and social studies. In 1992-93, arts & humanities and practical living/vocational studies were added to the performance events testing. In 1994-95, a maximum of 3 performance events were administered in each content area. Each student tested completes a single performance event. The number of events administered in each content area is presented in Table 1.

Table 1: KIRIS 1994-95: Number of Performance Events Administered by Content Area

	Mathematics	Science	Social Studies	Arts & Humanities	Practical Living
Grade 4	2	2	2	3	3
Grade 8	2	1	2	3	3
Grade 11	1[2]	2	2	3	2

After testing each year, all performance events are released to the public. Because of the nature of the performance events, it is not feasible to attempt to maintain the security of the tasks and reuse a performance event in a subsequent test administration. Therefore, unlike the open-response portion of the assessment in which a samples of items remain secure and are reused, results from successive administrations performance events testing cannot be linked through common item equating.

---

[1] Students at grade 11 first participated in the performance event testing in 1994-95. In prior years, grade 12 students participated in the performance event testing.

[2] Two performance events were administered and scored, but one event was dropped because of administration errors.

Prior to testing in 1992-93 a number of models were considered to link results from the 1992-93 performance events to the performance level scales established in 1991-92. The final two options considered involved common person equating by administering performance events from the base year and the current testing year

- to Kentucky students in non-accountability grades (e.g. students at grades 5, 9, or 11), or
- to out-of-state students in the accountability grades.

After consideration of the advantages and disadvantages of both approaches, the decision was made to select the out-of-state model primarily because of the widespread distribution and use of the released performance events in schools throughout Kentucky.

This equating model was used to link results from the 1992-93 and 1993-94 performance events to the 1991-92 scale. To establish performance level cutpoints for the 1992-93 and 1993-94 performance events, out-of-state performance event equating studies were conducted each year. For the 1992-93 administration, all performance events from 1991-92 and 1992-93 were administered to a sample of students in New Hampshire. Each student completed a single event from each year. For the 1993-94 administration, all performance events from the base year (1991-92 for mathematics, science, and social studies; 1992-93 for arts & humanities and practical living) and 1993-94 were administered to a sample of students in Maine. Once again, each student completed a single event from the current testing year and the base year. Descriptions of these studies can be found in the appropriate KIRIS technical reports: *KIRIS: 1992-93 Technical Report (KDE: Jan, 1993), KIRIS Accountability Cycle I Technical Manual (KDE: July, 1995).*

As Kentucky students gained experience with performance events, the appropriateness of continuing to use out-of-state sample of students to equate results was questioned. This concern, in addition to a) the logistical difficulties in obtaining an adequate equating sample and b) perceptions of problems with using an out-of-state sample led to the decision by the Kentucky Department of Education (KDE) to select an alternative method to link the 1994-95 performance event results to the established scales.

One model explored was a data-based approach to linking the performance event results to the proficiency level scales by linking them through the open-response items. Results from the performance events would be scaled with results from the open-response tests to create a direct link to the performance level scales established for the open-response tests. Data presented at the December, 1995 meeting of the *KIRIS National Technical Working Group* (NTWG) raised concerns about the validity of this approach. In short, neither student level nor school level correlations between performance on open-response items and performance events appeared to be strong enough to justify the scaling of performance events with open-response items.

An alternative model presented to the NTWG at the December, 1995 meeting was an expert judgment approach to linking the assessments. The basic premise of the model was that expert judges would determine performance level cutpoints for the 1994-95 performance events by comparing samples of student work from 1994-95 performance events to student work on the

1993-94 performance events. From discussions at that meeting, the general design of the 1994-95 performance event equating study evolved:

- Benchmark folders containing samples of 1993-94 student work at the proficiency level cutpoints would be compiled.
- Folders containing samples of 1994-95 student work at each performance event score point would be compiled.
- Scorers from Advanced Systems would serve as judges to categorize performance at each 1994-95 score point by making comparisons to the 1993-94 benchmark folders.

Following the NTWG meeting the final design and procedures for the equating study were developed with input from Advanced Systems (ASME), KDE, and Edward Haertel, a member of the NTWG.

Procedures and Results

The study contained two distinct phases. In Phase I, benchmark folders would be created containing samples of 1993-94 student work at each of the performance level cutpoints. In Phase II, judges would categorize 1994-95 student work at each score point as Novice, Apprentice, Proficient, or Distinguished by comparing it to the level of performance exhibited in the benchmark folders. Information about the judges is presented following the description of these two phases.

Phase I

The process used to create the benchmark folders was to select samples of student work at the raw score performance level cutpoints from each 1993-94 task. For example, the raw score to performance level conversions for a grade 8 social studies task were as follows:

Novice	Apprentice	Proficient	Distinguished
0-5	6-8	9-10	11+

Therefore, Novice/Apprentice benchmarks could be selected from students scoring 5 or 6, Apprentice/Proficient from students scoring 8 or 9, and Proficient/Distinguished from students scoring 10 or 11. (See Appendix 9-3 of the KIRIS Accountability Cycle I Technical Manual for raw score conversions associated with other performance events.)

This procedure was complicated, however, by the scoring system used to score 1993-94 performance events in Grade 4 reading, mathematics, and social studies. In those areas, all performance events were scored on a 0-4 scale. The limited number of score presented three challenges to the new equating process.

1. Because the four score points 1-4 were being mapped onto four performance levels, there were several cases where a performance level was represented by a single score point. This single score point, therefore, might serve three roles: benchmark for performance at one level, cutpoint for performance at the adjacent lower level, cutpoint for performance at the adjacent higher level.

2. The combination of a limited number of score points and the policy decision that perfect scores of 4 be assigned to the Distinguished level resulted in several performance events on which it was not possible to exhibit proficient performance. Performance was classified as Novice, Apprentice, or Distinguished.

3. The limited number of possible score points increased the difficulty of equating decisions in 1992-93 and 1993-94. The overall goal of the assignment of individual score points to performance levels was to achieve the desired performance level distributions across all tasks rather than on each individual task.

To overcome these challenges, seperate procedures were required to select the benchmark papers. One system was developed for grade 4 mathematics, science, and social studies. A second system was employed for the remaining content areas.

Grade 4 mathematics, science, and social studies

Given the limited number of score points available, it was assumed that a wide range of performance was exhibited within each score point. It was essential, therefore, that the process used to select benchmark papers account for this range of performance. The following seven step procedure was developed to identify benchmark papers.

Step 1: Select 9 student response papers at each score point 1, 2, 3, and 4 for each event.

For each performance event, 9 papers representing each score point were selected. The decision to select an equal number of papers at each score point rather than a stratified sample based on the percentage of students receiving each score point was made because judges needed to review a large enough sample of work at each score point to determine the level of performance exhibited. In mathematics, therefore, where there are three performance events, a total of 108 papers were selected.

Grade 4 Mathematics: Number of papers selected at each score point.

Task	Score 1	Score 2	Score 3	Score 4
MA1	9	9	9	9
MA2	9	9	9	9
MA4	9	9	9	9

Step 2: Selected papers at each event/point were placed in a folder and clearly labeled.

Step 3: Judges rank ordered the nine papers in each folder from lowest to highest based on performance.

The papers in each folder were sorted by teams of judges. At least two judges reviewed papers at each score point. Judges in each team worked together to reach consensus on the order of papers within the folder. The judges did not review all of the tasks used to assign the original performance levels. Several of the performance events were classified as interdisciplinary tasks and the single 1-4 score was used in multiple content areas. Performance events were not included in a content area review if judges determined that the performance required by the task

a) was not representative of their content area and
b) would not be useful in classifying performance on the 1994-95 performance events.

There were 2 performance events reviewed in mathematics, 3 in science, and 3 in social studies. An example depicting the result this initial ranking is presented in the following table:

Task MA1 - ■	Task MA2 - ▲	Task MA4 - ◆
■ - 1L	▲-1L	◆-1L
■ - 1M	▲-1M	◆-1M
■ - 1H	▲-1H	◆-1H
■ - 2L	▲-2L	◆-2L
■ - 2M	▲-2M	◆-2M
■ - 2H	▲-2H	◆-2H
■ - 3L	▲-3L	◆-1L
■ - 3M	▲-3M	◆-1M
■ - 3H	▲-3H	◆-1H
■ - 4L	▲-4L	◆-4L
■ - 4M	▲-4M	◆-4M
■ - 4H	▲-4H	◆-4H

Step 4: The middle paper from the low, middle, and high groups in each folder were selected.

Three papers from each folder of 9 rank ordered papers representing the low (2), middle (5), and high (8) groups. These papers were labeled with task, score point, and designation (low, middle, high). The result was a set of 24 or 36 papers, dependent upon the number of events reviewed in each content area.

Step 5: Rank order the selected set of papers.

Judges rank ordered the set of representative papers in terms of level of performance. The entire pool of judges within the content area reviewed the papers as a group and reached consensus on the order of papers. The judges recorded their final rankings on worksheets.

Step 6: Apply proper weights to determine location of performance level cutpoints

After the papers were rank ordered, weights were assigned corresponding to the percentage of students statewide receiving each event/score combination. The percentage of students receiving each score point was divided equally among the low, middle, and high papers. Those weights were used to determine the benchmark papers representative of the performance level cutpoints.

For example, if 5% of the students were Distinguished on the 1993-94 mathematics performance events, the goal was to identify the paper that cut off the top 5% of the distribution. Adjacent papers were also included in the selection of the final set of papers to represent the cutpoints.

Step 7: Select papers to represent the cutpoints

Sample papers at each of the selected and adjacent levels were selected from the ordered folders of event/score point papers that were created in Step 3. These papers formed the anchor set describing work at the performance level cutpoint. Therefore, there was a set of 27 papers selected in each content area - 9 each at the N/A, A/P, and P/D cutpoints.

Grade 8 and 12 (All Content Areas) and Grade 4 Arts & Humanities, Practical Living

Because most performance events in these content areas were scored on scales with more than 4 score points, it was not necessary to select samples of papers from all score points. The only score points sampled were those adjacent to performance level cutpoints. For example, on grade 12 social studies event SS2, the score point to performance level assignments were:

	Novice	Apprentice	Proficient	Distinguished
Grade 12 - SS2	0-8	9-11	12-14	>14

Therefore, the score points adjacent to the cutpoints are N/A: 8,9, A/P: 11,12, and P/D: 14,15. These six score points were sampled. The complete five step process used to select the benchmark papers follows:

Step 1: Identify score points adjacent to performance level cutpoints

Step 2: Select 9 papers at each score point to be sampled.

Step 3: Selected papers were placed in a folder and clearly labeled.

Step 4: Judges rank ordered the papers in each folder from lowest to highest based on level of performance.

The papers in each folder were sorted by a team of two judges. At least two judges reviewed each folder. Judges worked together to reach consensus on the order of papers within the folder.

Step 5: Select 2 papers from each pair of performance level cutpoints.

Two papers from each pair of performance level cutpoints were selected. For example, if the N/A cutpoint was represented by the score points 8 and 9, the highest ranked 8 and the lowest ranked 9 would be selected as the papers representative of the cutpoint for that task.

Papers selected at each score point: Grade 12 SS2

	N/A -8	N/A-9	A/P-11	A/P-12	P/D-14	P/D-15
SS2	Highest	Lowest	Highest	Lowest	Highest	Lowest

If there were 4 tasks in a content area, eight papers were selected at each cutpoint (2 per task) or a total of twenty-four papers across the three cutpoints: N/A, A/P, and P/D. These papers served as the anchor sets describing work at the performance level cutpoints.

Phase II

The cutpoints for the 1994-95 performance events were determined by comparing performance at each score point to the 1993-94 anchor points. For each score point, judges determined whether the performance exhibited on a set of papers falls at, above, or below the 1993-94 anchors. These judgments were aggregated to determine the 1994-95 performance level cutpoints on each item. The following seven step procedure was conducted for all content area/grade level combinations.

Step 1: Select 5 papers at each score point

A set of 5 papers was selected representing each possible score point. The number of papers selected in this phase was reduced from 9 to 5 because of differences in the judgments required at this phase and the increased number of score points available. It was decided that five papers should be a sufficient sample because

a) there were a large number of score points (making finer scoring distinctions possible), and
b) papers were selected from all score points

The number of possible score points in a content area/grade level ranged from 12 in some content areas to as many as 36 in others.

Step 2: Selected papers were placed in a folder and clearly labeled.

Step 3: Judges reviewed the anchor set folders representing the performance level cutpoints.

Before reviewing any 1994-95 papers, each judge reviewed all of the papers in the benchmark folders representing the performance level cutpoints.

Step 4: Judges compared the performance exhibited by papers in each folder to the performance exhibited in the 93-94 anchor sets.

Each judge compared all folders with the 93-94 standards. In some content areas this was as many as 36 folders (for a total of 180 papers). A single rating was made for each folder, placing it in one of four performance level categories. Novice, Apprentice, Proficient, Distinguished.

It is important to note that each judge made an overall ranking for the entire folder. Judges did not make a fine distinction about individual papers within the folder. The overall impression of the work in the folder was determined.

Judges worked individually at this task. There were at least 14 judges per content area. Judges indicated their rating of each folder on a score sheet, and all score sheets were collected.

Step 5: Collect score sheets

Step 6: Aggregate score sheet data across judges

The ratings of judges on each score point were tallied. That is, the number of judges rating each score point Novice, Apprentice, Proficient, or Distinguished were determined.

Step 7: Determine performance level cutpoints for each task

Score point to performance level mappings were made for each score point. The basic rule followed was that the performance level chosen by a plurality of the judges was assigned to the score point. In the event of a tie, the higher performance level was assigned to the score point. One exception to the assignment rule was applied in the case of internal inconsistencies in ratings. In the rare cases where a score point was mapped to a lower performance level than a preceding score point, that score point was assigned to the same level as the preceding score point. Tallies and performance level assignments are presented at the conclusion of the text to this document.

There were two irregularities to note from the process. First, on grade 8 mathematics performance event MA-2 there was not a sufficient number of papers available to make judgments on two score points. Score points 15 and 16 were achieved by a total of 11 students statewide. These score points were assigned to the Distinguished level. Second, training leaders in Arts & Humanities may have felt that a cluster of folders on a grade 8 performance event contained a large number of misscored papers. The training leaders were asked to make classification judgments for these folders. The classifications of the training leaders confirmed the judges decisions so no adjustments were necessary.

Judges

Judges participating in this study were experienced Advanced Systems scorers. The primary criteria for recruiting judges for the study were prior experience scoring KIRIS performance events and overall scoring experience. All selected scorers met established guidelines for scoring KIRIS assessment items. A summary of scorer background characteristics by content area are provided below.

Background characteristics of scorers

Content	# of scorers	Median Education Level	Median Scoring Experience (hours)	% of scorers with performance event experience
Mathematics	15	BA/BS	803	100
Science	14	BA/BS	1656	93
Social Studies	14	BA/BS	2716	100
Arts & Humanities	16	BA/BS	2338	100
Practical Living	20	BA/BS	2812	100

Study activities within each content area were coordinated by two experienced Advanced Systems' scoring training leaders. In addition to clerical tasks related to the distribution of folders and score sheets, the training leaders explained the study procedures, reviewed performance event scoring guides, answered scorer questions, and monitored activities during the study.

Training

Training for the judges was conducted in two phases. Training leaders received initial training before the beginning of the study. Judges received training and instruction at the beginning of each daily session.

Training leaders were provided with a description of the purpose, design, and procedures of the study in a 2-hour training session conducted three days before the beginning of the study. Step-by-step descriptions of the procedures, similar to those included in this report, were provided to the training leaders for their review.

Each daily session began with a review of procedures for the training leaders and training session for judges. Training leaders explained the procedures to the judges, answered questions, and closely monitored the initial activities of the judges. These training sessions lasted approximately one half hour.

Schedule

The study was conducted in four sessions over nine days from December 19, 1995 to December 27, 1995. Phase I was conducted for all grade levels on the first day of the study (December 19). In each content are, judges were divided among the three grade levels. Phase II was conducted over the remaining three days (December 21, December 22, December 27). A single grade level was completed each day. All judges participated at all grade levels in Phase II.

Discussion

At the beginning of Phase II, judges were quite anxious about making the required comparative judgments. The major cause of their concern was the variety of skills measured by performance events within a content area. Comments about comparing 'apples and oranges' were frequent on the first day. As judges were directed to focus on the level of performance required in general terms rather than the specific content of a performance event, their concerns appeared to lessen. By the second day of Phase II, judges seemed comfortable with their tasks and proceeded through the activity with little difficulty.

1. Overall, there was a high level of consistency among ratings. As expected, judges' ratings were highly internally consistent. There was also a fairly high degree of consistency across judges. Evidently, judges quickly internalized a common perception of Novice, Apprentice, Proficient, and Distinguished performance. What is not known from this study, however, is whether judges performance level standards are grounded on the performance exhibited in the benchmark folders or on judges prior knowledge of the performance levels. This should not be a problem if there is a match between the level of performance in the benchmark folders and performance commonly accepted as Novice, Proficient, etc.

However, because original performance event standard setting was a statistical process rather than a judgment-based process, it cannot be assumed that this is the case. Future judgment-based equating studies could attempt to determine the accuracy of judges' standards by including a task from the base year in the judgment process.

RESULTS

Grade 11/12 Mathematics

MA-2	Novice	Apprentice	Proficient	Distinguished
1N	11	1	-	-
2N	10	2	-	-
3N	10	2	-	-
4N	9	3	-	-
5N	11	1	-	-
6A	6	6	-	-
7A	6	6	-	-
8A	1	11	-	-
9A	1	8	3	-
10P	-	6	6	-
11P	-	6	6	-
12P	-	3	9	-
13P	-	4	8	-
14P	-	3	9	-
15P	-	1	11	-
16P	-	-	9	3
17P	-	-	9	3
18P	-	-	9	3
19D	-	-	5	7
20D	-	-	-	-

Grade 11/12 Social Studies

SS-1	Novice	Apprentice	Proficient	Distinguished
1N	13	-	-	-
2N	13	-	-	-
3N	9	4	-	-
4A	6	7	-	-
5A	3	10	-	-
6A	1	9	3	-
7P	1	6	6	-
8P	-	5	8	-
9P	-	8	5	-
10P	-	4	9	-
11P	-	-	10	3
12D	-	-	2	11

SS-1	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	9	3	-	-
3A	2	10	-	-
4A	-	8	4	-
5A	1	7	4	-
6P	-	5	7	-
7D	-	-	5	7
8D	-	-	1	11

Grade 11/12 Science

SC-1	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	12	-	-	-
3N	12	-	-	-
4N	12	-	-	-
5A	4	8	-	-
6A	8	4	-	-
7A	1	11	1	-
8A	2	11	-	-
9A	-	12	1	-
10A	1	8	3	-
11P	-	5	6	-
12P	-	4	8	-
13P	-	1	11	-
14P	-	1	11	-
15P	-	-	12	-
16D	-	-	6	6

SC-2	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	7	6	-	-
3A	5	7	-	-
4A	4	9	-	-
5A	7	6	-	-
6A	5	8	-	-
7A	2	11	-	-
8A	1	12	-	-
9A	1	12	-	-
10A	1	11	1	-
11A	-	8	4	-
12P	-	1	11	-
13P	-	2	10	-
14D	-	-	4	8
15D	-	-	-	-
16D	-	-	-	-

Grade 11/12 Practical Living

PL-1	Novice	Apprentice	Proficient	Distinguished
1N	16	1	-	-
2N	16	1	-	-
3A	6	11	-	-
4A	1	16	-	-
5A	3	13	1	-
6A	1	13	2	-
7A	-	16	1	-
8P	-	7	10	-
9P	-	1	16	-
10D	-	-	9	8

PL-2	Novice	Apprentice	Proficient	Distinguished
1N	17	-	-	-
2N	15	2	-	-
3N	14	3	-	-
4N	15	2	-	-
5A	5	12	-	-
6A	1	16	-	-
7A	2	15	-	-
8A	-	11	6	-
9A	-	11	6	-
10P	-	8	9	-
11P	-	-	12	5
12D	-	-	4	13
13P	-	-	10	7

Grade 11/12 Arts & Humanities

AH-1	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	12	-	-	-
3A	2	10	-	-
4A	-	12	-	-
5A	1	9	2	-
6P	-	1	11	-
7P	-	1	9	2
8D	-	1	4	7

AH-2	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	12	-	-	-
3N	7	5	-	-
4A	-	11	1	-
5A	-	10	2	-
6A	1	11	-	-
7P	-	-	7	5
8D	-	-	6	6

AH-3	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	10	2	-	-
3N	8	4	-	-
4A	-	12	-	-
5P	-	6	6	-
6P	-	2	10	-
7D	-	-	4	8
8D	-	-	8	4

Grade 8 Mathematics

MA-1	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	12	-	-	-
3N	12	-	-	-
4N	10	2	-	-
5A	5	7	-	-
6A	1	11	-	-
7A	-	11	1	-
8A	-	8	4	-
9P	-	3	8	1
10D			6	6
11D	-	-	-	-
12D	-	-	-	-

MA-2	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	12	-	-	-
3N	12	-	-	-
4N	12	-	-	-
5N	11	1	-	-
6N	7	5	-	-
7A	1	11	-	-
8A	1	11	-	-
9A	1	10	1	-
10A	-	8	4	-
11A	-	8	4	-
12P	-	6	6	-
13P	-	1	10	1
14P	-	2	10	-
15D	-	-	-	-
16D	-	-	-	-
17D	-	1	3	8
18D	-	-	-	-

Grade 8 Science

SC-2	Novice	Apprentice	Proficient	Distinguished
1N	11	-	-	-
2N	11	-	-	-
3N	10	1	-	-
4N	11	-	-	-
5A	9	2	-	-
6A	5	6	-	-
7A	-	11	-	-
8A	1	9	1	-
9P	-	11	-	-
10D	-	9	2	-
11D	-	4	7	-
12	-	4	7	-
13	-	3	8	-
14D	2	1	7	1

Grade 8 Social Studies

SS-1	Novice	Apprentice	Proficient	Distinguished
1N	13	-	-	-
2N	13	-	-	-
3N	13	-	-	-
4N	13	-	-	-
5N	10	3	-	-
6N	3	10	-	-
7A	1	12	-	-
8A	1	12	-	-
9A	-	13	-	-
10A	-	11	2	-
11A	-	9	4	-
12P	-	1	12	-
13P	-	-	13	-
14P	-	1	12	-
15D	-	-	3	10
16D	-	-	-	13

SS-2	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	12	-	-	-
3N	12	-	-	-
4N	11	1	-	-
5A	6	6	-	-
6A	3	9	-	-
7A	1	11	-	-
8A	-	11	1	-
9P	-	4	8	-
10D		4	8	-
11D	-	-	10	2
12D	-	-	1	11

Grade 8 Arts & Humanities

AH-1	Novice	Apprentice	Proficient	Distinguished
1N	7	-	-	-
2N	7	-	-	-
3N	5	2	-	-
4A	-	7	-	-
5A	1	4	2	-
6A	1	6	-	-
7A	-	4	2	1
8A	-	4	2	1

AH-2	Novice	Apprentice	Proficient	Distinguished
1N	7	-	-	-
2N	4	3	-	-
3N	4	3	-	-
4A	1	6	-	-
5A	1	6	-	-
6P	-	3	4	1
7P	-	2	3	2
8D	-	-	2	5

AH-3	Novice	Apprentice	Proficient	Distinguished
1N	7	-	-	-
2N	6	1	-	-
3A	-	7	-	-
4A	1	6	-	-
5A	-	4	3	-
6P	-	2	5	-
7P	-	-	5	2
8D	-	-	2	5

Grade 8 Practical Living

PL-1	Novice	Apprentice	Proficient	Distinguished
1N	15	1	-	-
2N	14	2	-	-
3N	15	1	-	-
4A	5	11	-	-
5A	-	16	-	-
6A	1	15	-	-
7A	-	10	6	-
8P	-	2	14	-
9P	-	-	16	-
10P	-	-	16	-
11D	-	1	5	10
12D	-	1	12	3

PL-2	Novice	Apprentice	Proficient	Distinguished
1N	16	-	-	-
2N	14	2	-	-
3N	12	4	-	-
4A	7	9	-	-
5A	-	15	1	-
6P	-	5	11	
7P	-	1	15	-
8D	-	-	3	13

PL-3	Novice	Apprentice	Proficient	Distinguished
1N	16	-	-	-
2N	16	-	-	-
3N	14	1	1	-
4A	1	14	1	-
5A	-	16	-	-
6A	-	11	5	-
7P	-	1	15	-
8P	-	1	15	-
9D	-	1	4	11

Grade 4 Practical Living

PL-1	Novice	Apprentice	Proficient	Distinguished
1N	14	-	-	-
2N	13	1	-	-
3N	6	8	-	-
4A	3	11	-	-
5A	-	13	1	-
6A	1	4	9	-
7A	2	10	3	-
8P	-	1	12	-
9P	-	-	1	13

PL-2	Novice	Apprentice	Proficient	Distinguished
1N	13	1	-	-
2N	13	1	-	-
3N	3	11	-	-
4A	-	14	-	-
5A	-	3	11	-
6P	-	2	9	3

PL-3	Novice	Apprentice	Proficient	Distinguished
1N	14	-	-	-
2N	13	1	-	-
3A	5	9	-	-
4A	-	14	-	-
5A	2	12	-	-
6A	-	8	6	-
7P	-	4	10	-
8P	-	1	13	-
9D	-	-	-	14

Grade 4 Arts & Humanities

AH-1	Novice	Apprentice	Proficient	Distinguished
1N	11	-	-	-
2N	11	-	-	-
3N	10	-	-	-
4N	9	2	-	-
5N	6	5	-	-
6A	2	9	-	-
7A	-	9	2	-
8A	-	10	1	-
9A	-	6	5	-
10P	-	3	8	-
11P	-	2	6	3
12D	-	2	1	8

AH-2	Novice	Apprentice	Proficient	Distinguished
1N	13	-	-	-
2N	11	2	-	-
3A	1	12	-	-
4A	-	13	-	-
5A	-	9	4	-
6P	-	4	9	-
7P	1	2	6	3
8D	-	1	5	6

AH-3	Novice	Apprentice	Proficient	Distinguished
1N	15	1	1	-
2N	17	-	-	-
3N	9	3	-	-
4N	8	4	-	-
5A	6	6	-	-
6A	3	9	-	-
7A	4	8	-	-
8A	1	12	-	-
9A	-	10	2	-
10P	-	6	6	-
11P	1	4	7	-
12P	-	1	11	-
13P	-	4	6	2
14P	-	7	2	3
15D	-	1	5	6
16D	-	1	7	5

Grade 4 Science

SC-1	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	12	-	-	-
3N	9	3	-	-
4A	6	6	-	-
5A	2	10	-	-
6A	1	9	2	-
7A	1	9	2	-
8A	2	9	1	-
9A	-	9	3	-
10A	-	9	2	1
11A	-	8	4	-
12D	-	2	10	-

SC-2	Novice	Apprentice	Proficient	Distinguished
1N	12	-	-	-
2N	12	-	-	-
3N	8	4	-	-
4A	2	10	-	-
5A	-	12	-	-
6A	-	8	4	-
7P	-	6	6	-
8P	-	4	8	-
9P	-	-	12	-
10D	-	-	11	1

Grade 4 Social Studies

SS-1	Novice	Apprentice	Proficient	Distinguished
1N	8	-	-	-
2A	3	5	-	-
3A	-	8	-	-
4P	-	2	6	-
5P	-	-	8	-
6D	-	-	-	8

SS-2	Novice	Apprentice	Proficient	Distinguished
1N	8	-	-	-
2N	-	8	-	-
3N	-	8	-	-
4A	-	-	8	-
5A	-	-	3	5
6A	-	-	-	8

Grade 4 Mathematics

MA-1	Novice	Apprentice	Proficient	Distinguished
1N	10	-	-	-
2N	10	-	-	-
3A	4	6	-	-
4A	2	8	-	-
5A	-	10	-	-
6A	-	8	2	-
7A	-	8	2	-
8P	-	2	8	-
9D	-	1	5	4

MA-2	Novice	Apprentice	Proficient	Distinguished
1N	10	-	-	-
2N	10	-	-	-
3A	4	6	-	-
4A	-	10	-	-
5A	-	8	2	-
6A	-	6	4	-
7P	-	-	8	2
8D	-	-	-	10

Appendix 10-1
Individual Student Report

GRADE 8: INTRODUCTION TO THE 1995–96 KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM (KIRIS) INDIVIDUAL STUDENT REPORT

This report tells how well you did on parts of the Kentucky Instructional Results Information System (KIRIS) Assessment in 1995-96. If you took the ***KIRIS On-Demand test***, it shows your performance in four subjects: reading, mathematics, science, and social studies. If you completed a writing and/or mathematics portfolio, your performance for each is included. For each portion of the test and/or portfolio, your ***performance level*** is reported as ***Novice, Apprentice, Proficient***, or ***Distinguished***. The Kentucky Board of Education's goal is for all students to reach at least the Proficient level.

I. KIRIS On-Demand Assessment — Open-Response Items

A. *Kentucky Percentile Rank* — This report gives your Kentucky Percentile Rank in each subject, shown as a diamond on the graph for reading, mathematics, science, and social studies. Percentile Rank tells how you performed compared to other Kentucky students who participated in the assessment in the spring of 1996. While it is interesting to know how you scored in relation to other students in Kentucky, your performance against specific standards is more important. You and your parent(s)/guardian(s) are encouraged to discuss your work in more detail with your teacher, who can show you examples of work at each level: Novice, Apprentice, Proficient, and Distinguished.

B. *Score Range (Graphically displayed as a bar around student Kentucky Percentile Ranks)* — You might think about your test score(s) as a range. If you had taken a different version of the test or had taken the same test on another day, you probably would not have received exactly the same score. The bars surrounding the Percentile Ranks in the graph on the other side of this report show the range within which you would probably score. Your score range gives a more complete picture of your score possibilities. Educators know this, and in fact, specifically ask that score ranges be included with scores.

C. *Performance Levels* — Open-response answers were scored by specially trained people who used detailed ***scoring guides*** for each of the questions answered. Based on careful reading of the work, they assigned score points to each answer. These score points were combined to determine the Performance Level for each subject: Novice, Apprentice, Proficient, or Distinguished.

Each year, the Department of Education releases some of the open-response test items and scoring guides. If you would like to view or obtain copies of these, please contact your school principal/counselor.

II. *Portfolios*

A. *Performance Levels* — Portfolios were scored by staff in your school who used a special scoring guide. This guide is very similar to the self-assessment/conference form available for students and teachers to use throughout the school year. If you are not familiar with this form, the criteria for scoring portfolios were also described in the brochures students should have received at the beginning of the school year. If you would like a copy of these brochures, please contact your school principal/counselor.

B. Scores — Portfolios are assigned a performance level: Novice, Apprentice, Proficient, or Distinguished. The percent of Kentucky students scoring at each performance level during the spring of 1996 is reported in the table on the reverse side of this page so that students and parents/guardians can compare individual student performance to a state reference point.

Note: Please refer to the Glossary for definitions of ***boldfaced italicized*** words. A copy of the Glossary may be obtained from your school principal/counselor.

1995–96 KENTUCKY INSTRUCTIONAL RESULTS INFORMATION SYSTEM (KIRIS) INDIVIDUAL STUDENT REPORT

Name: XXXXXXXXXX XXXXXXXXXXXXXX District: KIRIS School: KIRIS MIDDLE SCHOOL , Grade: 08

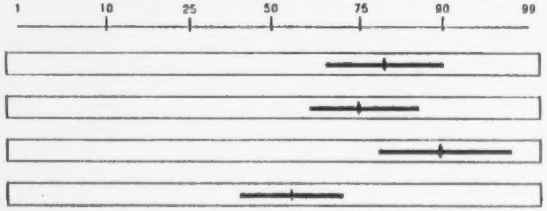
Subject	Your Performance Level	Percent of Kentucky Students in each Performance Level			
		Novice	Apprentice	Proficient	Distinguished
Reading	Apprentice	6	81	13	0 *
Mathematics	Proficient	36	38	15	12
Science	Apprentice	55	45	0 *	0 *
Social Studies	Apprentice	30	57	12	1
Writing Portfolio	Proficient	50	38	10	1
Mathematics Portfolio	Apprentice	55	34	9	1

The percent of Kentucky students across performance levels for each subject may not add to 100% due to rounding.

* Less than .5% scored at this level.

Subject	Your Kentucky Percentile Rank
Reading	80
Mathematics	75
Science	90
Social Studies	55

RANK: Percent of Kentucky Students at or Below Your Scores



You might think about your score(s) as a range. If you had taken the same or a different version of the test on another day, your scores would likely vary. Most of the time, your scores will fall somewhat above or below the percentile rank reported. The bars surrounding the percentile ranks on this graph above represent this range.

Description of Results:

READING:	Your Performance Level is Apprentice. You scored in the upper range of this performance level. Your Kentucky Percentile rank of 80 indicates that 80% of Kentucky students who took this test in the spring 1996 scored at or below your score.
MATHEMATICS:	Your Performance Level is Proficient. Your Kentucky Percentile rank of 75 indicates that 75% of Kentucky students who took this test in the spring 1996 scored at or below your score.
SCIENCE:	Your Performance Level is Apprentice. You scored in the middle range of this performance level. Your Kentucky Percentile rank of 90 indicates that 90% of Kentucky students who took this test in the spring 1996 scored at or below your score.
SOCIAL STUDIES:	Your Performance Level is Apprentice. You scored in the middle range of this performance level. Your Kentucky Percentile rank of 55 indicates that 55% of Kentucky students who took this test in the spring 1996 scored at or below your score.
TOTAL SCORE:	Your TOTAL SCORE Percentile Rank is based on your combined performance on the KIRIS Transitional (On-Demand) Test in reading, mathematics, science, and social studies, which consists of open-response test questions. Writing and/or mathematics portfolio(s) do not affect this score. Your Kentucky Percentile Rank of 79 indicates that 79% of Kentucky students who took this test in the spring 1996 scored at or below your score. If you had taken this same assessment again or a different version, your Total Score Percentile Rank would probably have fallen between 71 and 85.
WRITING PORTFOLIO:	Your Performance Level is Proficient. You scored at the same performance level as did 10% of the other students in the spring of 1996. As indicated in the table above: 50% scored Novice, 38% scored Apprentice, 10% scored Proficient, 1% scored Distinguished.
MATHEMATICS PORTFOLIO:	Your Performance Level is Apprentice. You scored at the same performance level as did 34% of the other students in the spring of 1996. As indicated in the table above: 55% scored Novice, 34% scored Apprentice, 9% scored Proficient, 1% scored Distinguished.

XXX
0000645

Appendix 10-2
Student Listing Report

KIRIS 1995-96 STUDENT LISTING

Grade: 08
Code: xxx
District:
School:

Page: 1 OF 2
Accountable Students 18
Run Date 12/04/96

Student Name/ Computer ID	Type of Result	Reading Open Response	Mathematics Open Response	Science Open Response	Social Studies Open Response	Total Battery Open Response	Writing Portfolio	Writing Portfolio Instructional Analysis										Math. Portfolio	Mathematics Portfolio Core Concepts						
								1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7
X..............., X..........	Performance	A	A	N	A		N											N	1	1	2	1	1	1	1
0809541	KY Percentile	22	44	17	75	38																			
X..............., X..........	Performance	A	D	A	P		A											N	1	1	1	1	1	1	1
0809542	KY Percentile	37	88	62	98	88				✓				✓											
X..............., X..........	Performance	A	N	N	A		N											N	1	1	B	1	1	1	B
0809543	KY Percentile	14	21	12	33	16																			
X..............., X..........	Performance	A	P	N	A		A											A	2	2	2	2	1	2	2
0809544	KY Percentile	44	73	33	53	51		✓																	
X..............., X..........	Performance	A	P	A	A		P											A	2	1	2	1	2	2	2
0809545	KY Percentile	80	75	90	55	79		✓	✓	✓															
X..............., X..........	Performance	A	A	A	N		N											N	1	1	1	2	1	1	2
0809546	KY Percentile	13	50	57	12	25																			
X..............., X..........	Performance	A	A	A	A		A											N	1	1	1	1	1	1	1
0809547	KY Percentile	61	65	66	45	60					✓					✓									
X..............., X..........	Performance	P	D	A	P		P											P	3	2	3	2	2	3	3
0809548	KY Percentile	98	98	85	91	98		✓		✓		✓				✓									
X..............., X..........	Performance	A	A	N	N		A											N	1	1	1	1	1	1	2
0809549	KY Percentile	33	53	18	23	28				✓		✓													

Performance Levels

D = Distinguished
P = Proficient
I = Incomplete
NT = Not Tested
A = Apprentice
N = Novice
B = Blank
NA = Not Applicable
* = Student excluded from Accountability

Writing Portfolio Instructional Analysis

1 = Purpose
2 = Audience
3 = Voice/Tone
4 = Developing Ideas
5 = Supporting Details
6 = Organization
7 = Transitions
8 = Sentences
9 = Language
10 = Editing

✓ = Commendations

Mathematics Portfolio

Core Concepts
1 = Number
2 = Mathematical Procedures
3 = Space & Dimensionality
4 = Measurement
5 = Change
6 = Mathematical Structure
7 = Data

Degree of Understanding of Core Concepts
B = Blank
1 = Basic Understanding
2 = Understanding
3 = Broad Understanding
4 = Comprehensive Understanding With Connections
* = Double Marks
Blank = No Data Available

BEST COPY AVAILABLE

KIRIS 1995-96 STUDENT LISTING

Grade: 08
Code: XXX
District: /
School:

Page: 2 OF 2
Accountable Students: 18
Run Date: 12/04/96

Student Name/ Computer ID	Type of Result	Reading Open Response	Mathematics Open Response	Science Open Response	Social Studies Open Response	Total Battery Open Response	Writing Portfolio	Writing Portfolio Instructional Analysis										Math Portfolio	Mathematics Portfolio Core Concepts						
								1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	N	A	N	N		N											N	1	1	2	1	1	1	1
0809550	KY Percentile	3	68	23	10	14																			
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	A	A	N	N		N											N	1	1	1	1	1	1	1
0809521	KY Percentile	8	47	18	3	9				✓															
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	A	A	N	A		N											N	1	1	1	1	1	1	1
0809522	KY Percentile	22	54	19	57	35							✓												
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	A	N	N	N		P											N	1	1	1	1	1	1	1
0809523	KY Percentile	27	7	8	18	11		✓	✓	✓			✓												
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	N	N	N	N		N											N	1	1	1	1	1	1	1
0809528	KY Percentile	4	2	4	22	4																			
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	A	N	N	P		A											N	1	1	1	1	1	1	B
0809525	KY Percentile	52	29	48	88	59					✓				✓										
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	A	P	A	P		A											A	2	1	1	2	2	1	2
0809524	KY Percentile	77	73	90	88	86					✓		✓												
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	A	P	A	A		P											N	2	1	2	2	1	1	1
0809526	KY Percentile	54	75	58	69	67		✓	✓	✓		✓													
Xxxxxxxxxxxxxx, Xxxxxxxxx	Performance	A	P	N	P		A											A	2	2	2	2	1	2	2
0809527	KY Percentile	87	81	44	96	87		✓		✓															

Performance Levels

D = Distinguished
A = Apprentice
P = Proficient
N = Novice
I = Incomplete
B = Blank
NT = Not Tested
NA = Not Applicable
* = Student excluded from Accountability

Writing Portfolio Instructional Analysis

1 = Purpose
2 = Audience
3 = Voice/Tone
4 = Developing Ideas
5 = Supporting Details
6 = Organization
7 = Transitions
8 = Sentences
9 = Language
10 = Editing

✓ = Commendations

Mathematics Portfolio

Core Concepts
1 = Number
2 = Mathematical Procedures
3 = Space & Dimensionality
4 = Measurement
5 = Change
6 = Mathematical Structure
7 = Data

Degree of Understanding of Core Concepts
B = Blank
1 = Basic Understanding
2 = Understanding
3 = Broad Understanding
4 = Comprehensive Understanding With Connections
* = Double Marks
Blank = No Data Available

Appendix 10-3
Item Level Report

1995-96

Kentucky Instructional Results Information System

Grade 8 ITEM LEVEL REPORT FOR PERFORMANCE-BASED TASKS

(Open-ended items and writing on-demand)

Code: XXX
District: KIRIS
School: KIRIS MIDDLE SCHOOL

Page: 1 of 1
Group Size: 18
Test Date: April, 1996

			COMMON																												MATRIX SAMPLED													WRITING (On-Demand)	
			Reading							Math							Science							Social Studies							Form	Reading		Math		Science		Social Studies		A & H		PL / VS		Score	Task #
		Test Item #	1	2	3	4	5	6	PL	1	2	3	4	5	6	PL	1	2	3	4	5	6	PL	1	2	3	4	5	6	PL		1	2	1	2	1	2	1	2	1	2	1	2		
XXXXXXXXXXXX	XXXXXXX	0809521	2	2	1	1	1	2	A	1	1	1	0	4	2	A	2	1	0	1	1	0	N	0	0	0	0	0	2	N	06	0	1	1	1	0	0	0	0	1	2	1	2	N	11
XXXXXXXXXXXX	XXXXXXX	0809522	2	2	3	2	1	1	A	1	3	2	0	1	0	A	1	0	0	0	3	2	N	2	2	2	1	0	1	A	08	2	1	2	2	0	0	2	2	1	1	1	1	N	16
XXXXXXXXXXXX	XXXXXXX	0809523	3	2	2	0	2	2	A	1	0	0	1	1	0	N	1	2	0	0	0	0	N	0	2	0	0	0	2	N	01	0	2	0	0	0	0	1	1	1	1	2	2	N	01
XXXXXXXXXXXX	XXXXXXX	0809524	2	2	3	3	3	4	A	2	3	2	1	1	2	P	2	2	2	3	3	2	A	3	3	3	1	1	2	P	07	2	1	0	4	2	2	2	2	2	2	2	2	A	13
XXXXXXXXXXXX	XXXXXXX	0809525	3	2	2	3	1	3	A	1	3	0	1	1	0	N	2	2	1	2	1	1	N	3	3	3	2	2	2	P	07	2	1	0	1	1	0	1	1	2	2	2	2	N	13
XXXXXXXXXXXX	XXXXXXX	0809526	3	2	2	2	2	2	A	1	4	2	2	4	2	P	4	2	3	1	1	1	A	2	2	3	2	1	1	A	08	2	3	1	1	0	0	1	2	3	1	2	1	N	16
XXXXXXXXXXXX	XXXXXXX	0809527	4	2	3	3	3	2	A	1	2	2	3	2	3	P	1	0	0	3	2	1	N	2	2	2	3	3	3	P	09	1	3	2	3	1	2	3	3	3	2	3	2	N	18
XXXXXXXXXXXX	XXXXXXX	0809528	2	1	1	0	1	2	N	0	0	0	0	0	0	N	1	0	0	0	0	0	N	1	1	1	1	1	1	N	04	0	0	2	0	0	0	0	1	1	1	0	0	N	08
XXXXXXXXXXXX	XXXXXXX	0809541	1	1	2	3	0	2	A	2	1	0	1	4	1	A	0	1	0	2	2	0	N	2	2	2	2	2	1	A	09	2	2	0	3	0	0	2	2	2	1	2	2	N	17
XXXXXXXXXXXX	XXXXXXX	0809542	2	2	1	2	3	2	A	3	4	2	1	2	3	D	2	2	0	2	2	1	A	4	4	3	3	3	3	P	04	2	1	4	1	1	2	2	2	2	2	3	2	N	08
XXXXXXXXXXXX	XXXXXXX	0809543	1	2	1	2	2	1	A	1	2	0	0	1	1	N	3	1	0	0	0	0	N	2	1	2	0	2	1	A	02	1	1	0	0	1	0	1	1	0	1	1	2	N	03
XXXXXXXXXXXX	XXXXXXX	0809544	2	2	2	2	2	2	A	2	2	1	2	4	2	P	1	1	0	1	1	2	N	1	2	0	1	1	2	A	12	2	3	1	2	1	2	3	2	2	2	1	1	A	24
XXXXXXXXXXXX	XXXXXXX	0809545	4	4	3	2	2	3	A	2	1	1	3	4	3	P	2	2	2	3	3	3	A	2	1	1	2	1	1	A	12	2	2	1	2	1	2	3	2	2	2	1	2	A	24
XXXXXXXXXXXX	XXXXXXX	0809546	3	2	1	2	1	1	A	2	2	1	3	1	1	A	2	0	2	1	0	3	A	1	1	1	2	0	0	N	05	1	0	0	0	2	2	0	0	1	0	1	2	N	10
XXXXXXXXXXXX	XXXXXXX	0809547	2	2	3	2	2	2	A	2	1	2	2	1	3	A	1	2	0	2	2	2	A	2	2	2	2	0	1	A	11	2	3	2	0	2	2	0	1	1	1	2	1	N	22
XXXXXXXXXXXX	XXXXXXX	0809548	4	3	3	4	4	3	P	4	4	1	4	4	3	D	4	2	2	2	2	2	A	2	2	2	4	2	2	P	11	3	3	3	4	1	2	2	3	1	3	2	3	A	22
XXXXXXXXXXXX	XXXXXXX	0809549	3	3	1	2	2	2	A	1	4	1	0	1	2	A	1	0	0	0	1	0	N	0	1	2	1	0	1	N	10	0	1	1	1	1	2	0	2	1	1	2	2	N	20
XXXXXXXXXXXX	XXXXXXX	0809550	1	2	1	0	1	0	N	1	2	2	0	4	3	A	1	0	0	1	2	0	N	0	0	1	1	1	1	N	10	0	1	1	1	0	2	0	0	0	0	2	1	N	20
School Mean Score			2.4	2.1	1.9	1.9	1.8	2.0		1.6	2.2	1.1	1.3	2.2	1.7		1.7	1.1	0.7	1.3	1.4	1.1		1.6	1.7	1.7	1.6	1.1	1.5																
District Mean Score			2.4	2.1	1.9	1.9	1.8	2.0		1.6	2.2	1.1	1.3	2.2	1.7		1.7	1.1	0.7	1.3	1.4	1.1		1.6	1.7	1.7	1.6	1.1	1.5																
State Mean Score			2.4	2.2	2.1	2.2	1.9	2.3		1.9	1.6	1.5	1.3	1.1	1.3		1.7	1.2	1.3	1.5	1.7	1.1		1.7	1.5	1.4	1.4	1.2	1.5																

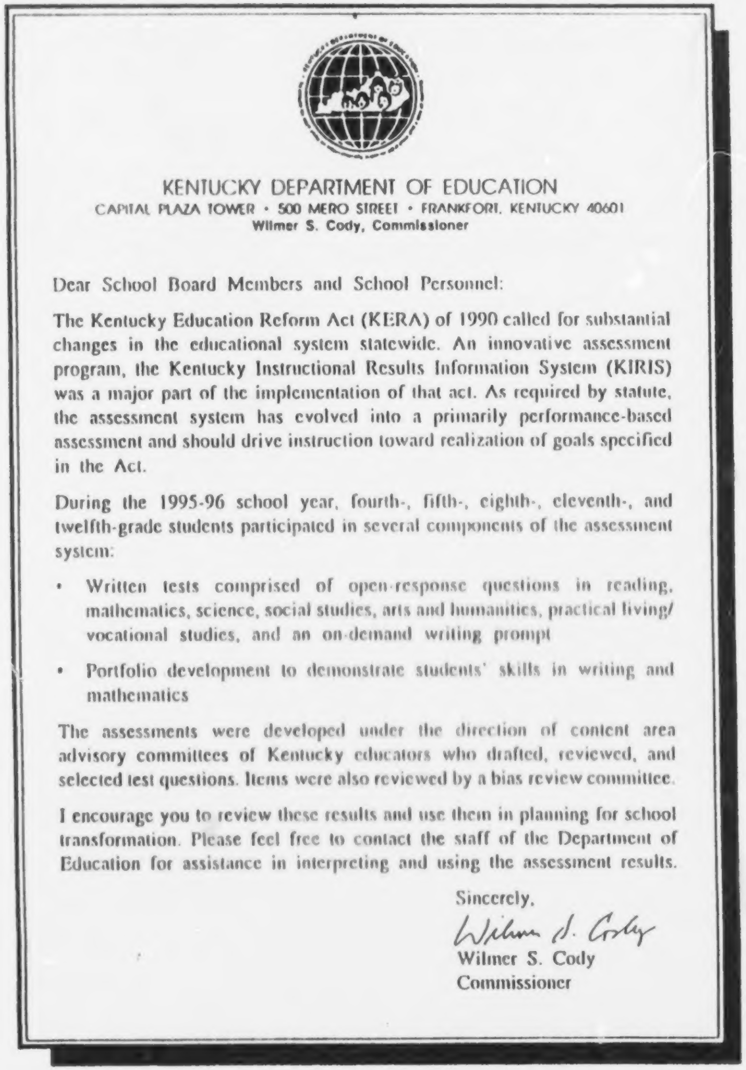
* Reassigned from non-A1 school.

Writing Task
NR = No response
OT = Off topic
NE = Non English

LEGEND: B = NO ATTEMPT BLANK = NO AVAILABLE DATA X = NO ITEM IN TEST FORM PL = PERFORMANCE LEVEL (Novice, Apprentice, Proficient, Distinguished) Run Date: October 11, 1996

BEST COPY AVAILABLE

Appendix 10-4
Assessment Curriculum Report



KENTUCKY DEPARTMENT OF EDUCATION
CAPITAL PLAZA TOWER • 500 MERO STREET • FRANKFORT, KENTUCKY 40601
Wilmer S. Cody, Commissioner

Dear School Board Members and School Personnel:

The Kentucky Education Reform Act (KERA) of 1990 called for substantial changes in the educational system statewide. An innovative assessment program, the Kentucky Instructional Results Information System (KIRIS) was a major part of the implementation of that act. As required by statute, the assessment system has evolved into a primarily performance-based assessment and should drive instruction toward realization of goals specified in the Act.

During the 1995-96 school year, fourth-, fifth-, eighth-, eleventh-, and twelfth-grade students participated in several components of the assessment system:

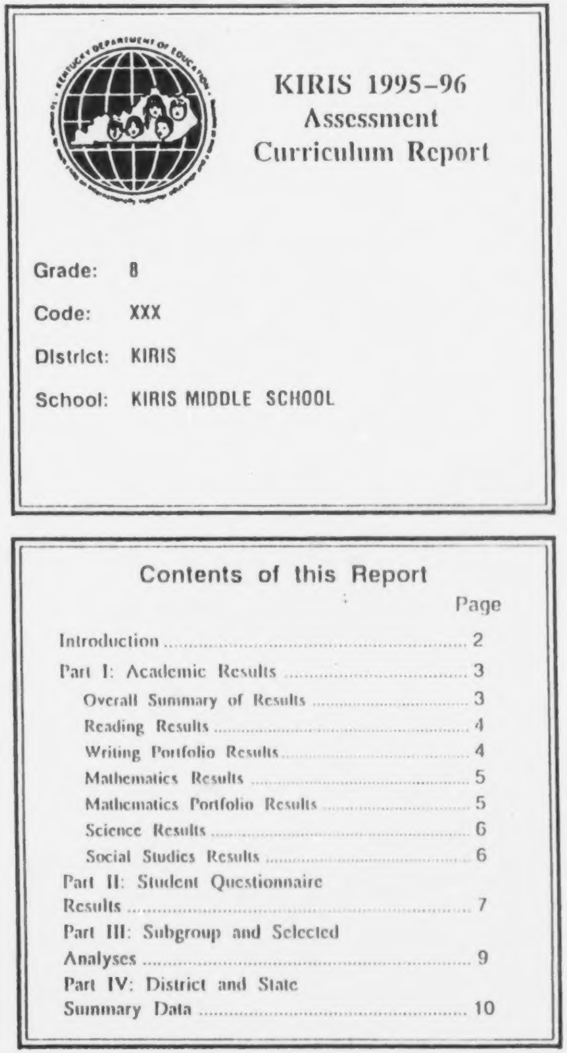
- Written tests comprised of open-response questions in reading, mathematics, science, social studies, arts and humanities, practical living/vocational studies, and an on-demand writing prompt
- Portfolio development to demonstrate students' skills in writing and mathematics

The assessments were developed under the direction of content area advisory committees of Kentucky educators who drafted, reviewed, and selected test questions. Items were also reviewed by a bias review committee.

I encourage you to review these results and use them in planning for school transformation. Please feel free to contact the staff of the Department of Education for assistance in interpreting and using the assessment results.

Sincerely,

Wilmer S. Cody
Wilmer S. Cody
Commissioner



KIRIS 1995–96
Assessment
Curriculum Report

Grade: 8

Code: XXX

District: KIRIS

School: KIRIS MIDDLE SCHOOL

Contents of this Report

Page
Introduction 2
Part I: Academic Results 3
  Overall Summary of Results 3
  Reading Results 4
  Writing Portfolio Results 4
  Mathematics Results 5
  Mathematics Portfolio Results 5
  Science Results 6
  Social Studies Results 6
Part II: Student Questionnaire Results 7
Part III: Subgroup and Selected Analyses 9
Part IV: District and State Summary Data 10

KIRIS 1995–96
Assessment Curriculum Report

Introduction

This student assessment curriculum report is part of a larger shipment of reports based on the 1995–96 Kentucky Instructional Results Information System (KIRIS) on-demand testing results for students in grades 4, 5, 8, 11, and 12. Other reports included in the shipment pertain to the performance of individual students. This student assessment curriculum report summarizes information for the school/district based on a summary of performance on the KIRIS On-Demand Assessment and the Writing and Mathematics Portfolio Assessments. These results also reflect performance of any students participating in the KIRIS Alternate Portfolio Assessment where such students were a part of the 4th, 8th, or 12th grade during 1995–96.

This report is divided into four parts: Part I details the summary of academic results; Part II provides a summary of student questionnaire items results; Part III provides results for various student subgroups; and Part IV provides summary performance level data for the district and state.

Students in grades 4, 8, and 11 completed batteries of open-response questions (referred to as KIRIS On-Demand Assessments) in reading, mathematics, science, and social studies. To provide broad coverage of each subject, 30 test questions at grades 8 and 11 (29 at grade 4) were developed for each content area and distributed across 12 test forms; each student completed only one form consisting of six common items and two matrix-sampled items for each area. Arts and humanities and practical living/vocational studies were included in the matrix portions of the tests.

At the 4th grade, each child took seven items in each content area: five common and two matrix-sampled. Each form contained two arts and humanities items and two practical living/vocational studies items. This approach, called "matrix-sampling," produces highly reliable results at school and district levels, while also reducing testing time for all students. The curriculum report is based on the entire battery of open-response questions.

Schools are held accountable for all of the students enrolled in the school as of the first day of the testing window.

KERA states that "schools shall expect a high level of achievement of all students." It also states that "schools shall be rewarded for an increased proportion of successful students, including those students who are at risk of school failure." Therefore, there are virtually no exclusions from the testing. Students not participating in on-demand tests were

- foreign exchange students;
- students with health-related conditions such that participation would threaten their health or well-being; and
- (at the school's option) limited English-speaking students who have been enrolled in an English-speaking school for fewer than two years.

The number and percent of students who did not participate for these reasons are provided on the bottom of page 3. Any other student for whom the school is accountable but who was not tested is assigned to the "Novice" performance level. The number and percentage of students not tested in on-demand testing, for whom a "Novice" rating is reported in the subject area reports, are also provided on page 3.

KIRIS 1995–96
Assessment Curriculum Report

Grade: 8
Code: XXX
District: KIRIS
School: KIRIS MIDDLE SCHOOL

PART I: Academic Results

Overall Summary of Results

Content Areas	Performance Level									
	Distinguished		Proficient		Apprentice		Novice		Nontested	
	#	%	#	%	#	%	#	%	#	%
Reading	0	0.00	1	5.56	15	83.33	2	11.11	0	0.00
Mathematics	2	11.11	5	27.78	7	38.89	4	22.22	0	0.00
Science	0	0.00	0	0.00	7	38.89	11	61.11	0	0.00
Social Studies	0	0.00	5	27.78	7	38.89	6	33.33	0	0.00
Arts and Humanities	0	0.00	3	8.33	12	33.33	21	58.33	0	0.00
Practical Living/Voc. Studies	0	0.00	0	0.00	3	8.33	33	91.67	0	0.00
Writing Portfolio	0	0.00	4	22.22	7	38.89	7	38.89	0	0.00
Mathematics Portfolio	0	0.00	1	5.56	4	22.22	13	72.22	0	0.00

NOTES:

The results in reading, mathematics, science, and social studies are based on the 30 (29 at grade 4) open-response (common and matrix) questions in each content area.

Results for arts and humanities and practical living/vocational studies are based on 24 matrix open-response questions in those two content areas.

In Curriculum Reports for the first Accountability Cycle, separate scores were computed for common and matrix items for each student. For this Curriculum Report, a single score was computed for each student based on both the common and matrix items, using an improved scaling procedure. To ensure a proper comparison with 1994–95 and 1995–96 data, scores were recomputed using this procedure for 1992–93 and 1993–94 results. Therefore, 1992–93 and 1993–94 data will differ from prior 1994-95 Curriculum Reports.

Alternate Portfolio	#	%
Students	0	0.00

Non-Participating Students*	#	%
Foreign exchange	0	0.00
Medical exemptions	0	0.00
Limited-English speakers	0	0.00

* for open-response on-demand testing

KIRIS 1995-96
Assessment Curriculum Report

Grade: 8
Code: XXX
District: KIRIS
School: KIRIS MIDDLE SCHOOL

Reading Results—On-Demand

	Distinguished		Proficient		Apprentice		Novice *	
	#	%	#	%	#	%	#	%
1995-96	0	0	1	6	15	83	2	11
1994-95	0	0	2	13	12	80	1	7
1993-94	0	0	2	13	10	67	3	20
1992-93	0	0	1	8	11	85	1	8

Last year you had 13.33% Proficient or Distinguished students; this year you had 5.56% Proficient or Distinguished

	Distinguished	Proficient	Apprentice	Novice *
SELECTION TYPE	%	%	%	%
Literary	0	11	78	11
Informational	0	11	89	0
Practical/Workplace	0	0	100	0
Persuasive	0	6	94	0

NOTE: Percentages across performance levels and selection type may not add to 100% due to rounding.

*Any nontested students have been assigned to the Novice category.

Writing Portfolio Results (developed over time)

	Distinguished		Proficient		Apprentice		Novice *	
	#	%	#	%	#	%	#	%
1995-96	0	0	4	22	7	39	7	39
1994-95	0	0	5	33	6	40	4	27
1993-94	1	7	2	14	6	43	5	36
1992-93	0	0	0	0	2	15	11	85

Last year you had 33.33% Proficient or Distinguished students; this year you had 22.22%.

Instructional Analysis**

	#	%
1 = Purpose	6	33
2 = Audience	3	17
3 = Voice/Tone	8	44
4 = Developing Ideas	3	17
5 = Supporting Details	3	17
6 = Organization	3	17
7 = Transitions	1	6
8 = Sentences	1	6
9 = Language	2	11
10 = Editing	0	0

**Student portfolios were holistically evaluated by teachers to determine the performance level. In addition, annotations were provided to further clarify the strengths of each portfolio. Above is a summary of the annotations given to students by their teachers.

KIRIS 1995–96
Assessment Curriculum Report

Grade: 8
Code: XXX
District: KIRIS
School: KIRIS MIDDLE SCHOOL

Mathematics Results—On-Demand

	Distinguished		Proficient		Apprentice		Novice *	
	#	%	#	%	#	%	#	%
1995-96	2	11	5	28	7	39	4	22
1994-95	1	7	1	7	8	53	5	33
1993-94	0	0	0	0	7	47	8	53
1992-93	0	0	1	8	4	31	8	62

Last year you had 13.34% Proficient or Distinguished students; this year you had 38.89% Proficient or Distinguished.

REPORTING CATEGORY	Distinguished %	Proficient %	Apprentice %	Novice * %
Numbers, Procedures & Structure	6	22	44	28
Space & Dimension, Measure	28	6	44	22
Change and Data	6	22	33	39

NOTE: Percentages across performance levels, reporting categories, and core concepts may not add to 100% due to rounding.

*Any nontested students have been assigned to the Novice category.

Mathematics Portfolio Results (developed over time)

	Distinguished		Proficient		Apprentice		Novice *	
	#	%	#	%	#	%	#	%
1995-96	0	0	1	6	4	22	13	72
1994-95	0	0	1	7	8	53	6	40
1993-94	0	0	6	40	7	47	2	13
1992-93	0	0	1*	8	7	54	5	38

Last year you had 6.67% Proficient or Distinguished students; this year you had 5.56% Proficient or Distinguished.

Mathematics Portfolio Core Concept Summary Data

	Blank		Basic Understanding		Understanding		Broad Understanding		Comprehensive Understanding	
CORE CONCEPTS	#	%	#	%	#	%	#	%	#	%
Number	0	0	12	67	5	28	1	6	0	0
Mathematical Procedures	0	0	15	83	3	17	0	0	0	0
Space & Dimensionality	1	6	10	56	6	33	1	6	0	0
Measurement	0	0	12	67	6	33	0	0	0	0
Change	0	0	15	83	3	17	0	0	0	0
Mathematical Structure	0	0	14	78	3	17	1	6	0	0
Data	2	11	9	50	6	33	1	6	0	0

KIRIS 1995-96
Assessment Curriculum Report

Grade: 8
Code: XXX
District: KIRIS
School: KIRIS MIDDLE SCHOOL

Science Results—On-Demand

	Distinguished		Proficient		Apprentice		Novice *	
	#	%	#	%	#	%	#	%
1995-96	0	0	0	0	7	39	11	61
1994-95	0	0	0	0	9	60	6	40
1993-94	0	0	0	0	3	20	12	80
1992-93	0	0	0	0	6	46	7	54

Last year you had 0.00% Proficient or Distinguished students; this year you had 0.00% Proficient or Distinguished.

	Distinguished	Proficient	Apprentice	Novice *
REPORTING CATEGORY	%	%	%	%
Nature of Scientific Activity, Models & Scales	0	0	44	56
Patterns, Constancy, and Evolution	0	0	22	78
Systems	0	0	39	61

NOTE: Percentages across performance levels and reporting categories may not add to 100% due to rounding.

*Any nontested students have been assigned to the Novice category.

Social Studies Results—On-Demand

	Distinguished		Proficient		Apprentice		Novice *	
	#	%	#	%	#	%	#	%
1995-96	0	0	5	28	7	39	6	33
1994-95	0	0	3	20	4	27	8	53
1993-94	0	0	1	7	11	73	3	20
1992-93	1	8	1	8	6	46	5	38

Last year you had 20.00% Proficient or Distinguished students; this year you had 27.78% Proficient or Distinguished.

	Distinguished	Proficient	Apprentice	Novice *
REPORTING CATEGORY	%	%	%	%
Democratic Principles and Political Systems	0	17	67	17
Social Systems, Economic Systems, and Cultural Diversity	0	11	67	22
Relation of Geography to Human Activity and Historical Perspective	0	17	50	33

BEST COPY AVAILABLE

Grade: 8
Code: XXX

PART II: Student Questionnaire Results

	#	%
How long have you lived in Kentucky?		
all my life	10	56
more than 5 years	8	44
from 1 to 5 years	0	0
less than 1 year	0	0
Did you go to kindergarten?		
yes	17	94
no	1	6
How many years have you attended this school?		
less than 1 whole year (I moved into this school partway through the year)	3	17
1 year	1	6
2 years	2	11
3 years	1	6
4 or more years	11	61
What kind of reader do you think you are?		
a poor reader	0	0
an average reader	7	39
a good reader	8	44
a very good reader	3	17
About how many books of your own, not counting your schoolbooks or comic books, do you have at home?		
5 or fewer	0	0
6-15	3	17
16-30	3	17
more than 30	11	61
How much television do you usually watch each day?		
none	0	0
1 hour or less	6	33
2-3 hours	4	22
4-5 hours	5	28
6 hours or more	3	17
How much time do you usually spend on homework each week?		
0-3 hours	6	33
4-7 hours	6	33
8-11 hours	5	28
12-15 hours	1	6
more than 15 hours	0	0
In a typical week, how often do your teachers give you specific, helpful feedback on your homework?		
every day	6	33
about 2 to 4 times a week	7	39
once a week	2	11
Never - I have no homework	0	0
Never - I have homework, but don't get this kind of feedback	3	17
How often are you read to by an adult?		
almost never	9	50
a few times a month	6	33
a few times a week	3	17
almost every day	0	0
How would you describe the test you have taken over the last few days? (Mark all that are true for you.)		
It made me think.	13	72
It was the hardest test I have ever taken.	1	6
It was the easiest test I have ever taken.	1	6
It was frustrating.	6	33
It was the most interesting test I have ever taken.	7	39

	#	%
How clear and understandable were the directions for the different tests?		
All the directions were very clear and understandable.	9	50
Some directions could have been clearer.	8	44
The directions were not very clear or understandable.	0	0
How many of the questions in the science portion of the test covered material that you learned in school in this or any earlier grade?		
Most of the questions covered material I learned in school.	10	56
Some of the questions covered material I learned in school.	6	33
None or very few of the questions covered material I learned in school.	2	11
How many of the questions on the social studies portion of the test covered material that you learned in school in this or any earlier grade?		
Most of the questions covered material I learned in school.	9	50
Some of the questions covered material I learned in school.	8	44
None or very few of the questions covered material I learned in school.	0	0
How many of the questions on the mathematics portion of the test covered material that you learned in school in this or any earlier grade?		
Most of the questions covered material I learned in school.	15	83
Some of the questions covered material I learned in school.	3	17
None or very few of the questions covered material I learned in school.	0	0
How many of the questions dealing with arts and humanities (art, music, dance or theater) covered material that you learned in school in this or any earlier grade?		
Most of the questions covered material I learned in school.	3	17
Some of the questions covered material I learned in school.	10	56
None or very few of the questions covered material I learned in school.	5	28
How many of the questions dealing with practical living skills (how we do things in every day life) or vocational studies (the kinds of things people do in their jobs) covered material that you learned in school in this or any earlier grade?		
Most of the questions covered material I learned in school.	6	33
Some of the questions covered material I learned in school.	9	50
None or very few of the questions covered material I learned in school.	2	11
Where did you get the calculator you used to complete the tests?		
I did not use a calculator.	0	0
I used my own calculator.	2	11
I used a calculator provided by the school.	16	89
If you did use a calculator, when did you use it?		
I used it only for the mathematics test.	8	44
I used it for the mathematics and science tests.	10	56
I used it for three or more tests.	0	0
I did not use a calculator	0	0
How much do you think your performance on these tests reflects how much you know?		
very much	9	50
somewhat	7	39
not very much	0	0
not at all, even though I tried my best	2	11
not at all, since I didn't even try	0	0

Student questionnaire results analyses reflect data as scanned from student answer documents. These analyses are based on tested students. Percentages across a question's response may not add to 100% due to information not provided by students or rounding.

BEST COPY AVAILABLE

PART II: Student Questionnaire Results (cont.)

How often do you do the following activities?

	2 or more times/week		about once a week		about once a month		never	
	#	%	#	%	#	%	#	%
working in groups	4	22	3	17	11	61	0	0
working on projects	1	6	3	17	14	78	0	0
looking up information	8	44	6	33	4	22	0	0
giving oral reports	1	6	2	11	14	78	1	6
using a calculator	14	78	3	17	1	6	0	0
reading books you choose	8	44	3	17	3	17	4	22
writing (journals, stories, other)	5	28	7	39	4	22	2	11
solving mathematics problems	15	83	2	11	1	6	0	0
using mathematics manipulatives	3	17	10	56	5	28	0	0

Student questionnaire results analyses reflect data as scanned from student answer documents. These analyses are based on tested students. Percentages across a question's response may not add to 100% due to information not provided by students or rounding.

PART III: Subgroup and Selected Analyses

	Male								Female							
	Distinguished		Proficient		Apprentice		Novice		Distinguished		Proficient		Apprentice		Novice	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Reading									0	0	1	9	10	91	0	0
Mathematics									2	18	5	45	2	18	2	18
Science									0	0	0	0	5	45	6	55
Social Studies									0	0	5	45	4	36	2	18
Arts & Humanities	0	0	0	0	2	14	12	86	0	0	3	14	10	45	9	41
Practical Living/Voc. St.	0	0	0	0	0	0	14	100	0	0	0	0	3	14	19	86
Writing Portfolio									0	0	4	36	6	55	1	9
Mathematics Portfolio									0	0	1	9	4	36	6	55

	White								Non-White							
	Distinguished		Proficient		Apprentice		Novice		Distinguished		Proficient		Apprentice		Novice	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Reading	0	0	1	6	15	83	2	11								
Mathematics	2	11	5	28	7	39	4	22								
Science	0	0	0	0	7	39	11	61								
Social Studies	0	0	5	28	7	39	6	33								
Arts & Humanities	0	0	3	8	12	33	21	58								
Practical Living/Voc. St.	0	0	0	0	3	8	33	92								
Writing Portfolio	0	0	4	22	7	39	7	39								
Mathematics Portfolio	0	0	1	6	4	22	13	72								

	Title 1 - (Targeted Assistance)								Non-Title 1 (Non-Targeted Assistance)							
	Distinguished		Proficient		Apprentice		Novice		Distinguished		Proficient		Apprentice		Novice	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Reading									0	0	1	6	15	83	2	11
Mathematics									2	11	5	28	7	39	4	22
Science									0	0	0	0	7	39	11	61
Social Studies									0	0	5	28	7	39	6	33
Arts & Humanities									0	0	3	8	12	33	21	58
Practical Living/Voc. St.									0	0	0	0	3	8	33	92
Writing Portfolio									0	0	4	22	7	39	7	39
Mathematics Portfolio									0	0	1	6	4	22	13	72

NOTE: Subgroup analyses reflect data as scanned from student answer documents. To protect anonymity, no data are reported if a category includes fewer than 10 students. Also, these analyses are based on tested students.

Analysis does not include Alternate Portfolios. Percentages may not add to 100% due to missing values or rounding.

Student counts in arts and humanities and practical living/vocational studies reflect the performance of students on each of two items in each of these academic areas. Thus, the student counts in these two areas may be up to double the counts found in the other academic areas.

BEST COPY AVAILABLE

KIRIS 1995–96
Assessment Curriculum Report

Grade: 8
Code: XXX
District: KIRIS
School: KIRIS MIDDLE SCHOOL

PART IV: District and State Summary Data

	Performance Level	Reading	Mathematics	Science	Social Studies	Arts and Humanities	Practical Living/ Vocational Studies	Writing Portfolio	Mathematics Portfolio
Percentage of Students in District at Each Performance Level	Distinguished	0	11	0	0	0	0	0	0
	Proficient	6	28	0	28	8	0	22	6
	Apprentice	83	39	39	39	33	8	39	22
	Novice	11	22	61	33	58	92	39	72
Percentage of Students in State at Each Performance Level	Distinguished	0	12	0	1	1	0	1	1
	Proficient	13	16	1	12	6	3	10	10
	Apprentice	81	36	45	57	31	15	38	34
	Novice	6	36	54	30	62	82	50	55

Percentages across performance levels and reporting categories may not add to 100% due to rounding. Percentages less than .5 have been rounded to 0.

Appendix 10-5
KIRIS Accountability Cycle II Report

KIRIS ACCOUNTABILITY CYCLE II

Grade: 08
Code: XXXXXX
District: KIRIS
School: KIRIS MIDDLE SCHOOL

Preliminary

	Academic Areas Performance Levels															
	% Distinguished				% Proficient				% Apprentice				% Novice			
	1992-93	1993-94	1994-95	1995-96	1992-93	1993-94	1994-95	1995-96	1992-93	1993-94	1994-95	1995-96	1992-93	1993-94	1994-95	1995-96
Reading																
Open Response	0.00	0.00	0.00	0.00	7.69	13.33	13.33	5.56	84.62	66.67	80.00	83.33	7.69	20.00	6.67	11.11
Mathematics																
Open Response	0.00	0.00	6.67	11.11	7.69	0.00	6.67	27.78	30.77	46.67	53.33	38.89	61.54	53.33	33.33	22.22
Performance Events																
Portfolio	0.00	0.00	0.00	0.00	7.69	40.00	6.67	5.56	53.85	46.67	53.33	22.22	38.46	13.33	40.00	72.22
Total	0.0000	0.0000	4.6690	7.7770	7.6900	12.0000	6.6700	21.1140	37.6940	46.6700	53.3300	33.8890	54.6160	41.3300	35.3310	37.2200
Science																
Open Response	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	46.15	20.00	60.00	38.89	53.85	80.00	40.00	61.11
Performance Events																
Total	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	46.1500	20.0000	60.0000	38.8900	53.8500	80.0000	40.0000	61.1100
Social Studies																
Open Response	7.69	0.00	0.00	0.00	7.69	6.67	20.00	27.78	46.15	73.33	26.67	38.89	38.46	20.00	53.33	33.33
Performance Events																
Total	7.6900	0.0000	0.0000	0.0000	7.6900	6.6700	20.0000	27.7800	46.1500	73.3300	26.6700	38.8900	38.4600	20.0000	53.3300	33.3300
Arts & Humanities																
Open Response	0.00	0.00	0.00	0.00	0.00	0.00	3.33	8.33	50.00	30.00	43.33	33.33	50.00	70.00	53.33	58.33
Performance Events																
Total	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	3.3300	8.3300	50.0000	30.0000	43.3300	33.3300	50.0000	70.0000	53.3300	58.3300
Prac. Living/Voc. Studies																
Open Response	0.00	0.00	3.33	0.00	0.00	0.00	0.00	0.00	16.67	35.71	50.00	8.33	83.33	64.29	46.67	91.67
Performance Events																
Total	0.0000	0.0000	3.3300	0.0000	0.0000	0.0000	0.0000	0.0000	16.6700	35.7100	50.0000	8.3300	83.3300	64.2900	46.6700	91.6700
Writing Portfolio	0.00	7.14	0.00	0.00	0.00	14.29	33.33	22.22	15.38	42.86	40.00	38.89	84.62	35.71	26.67	38.89

RUN DATE: 10/11/96

BEST COPY AVAILABLE

KIRIS ACCOUNTABILITY CYCLE II

Grade: 08
Code: XXXXXX
District: KIRIS
School: KIRIS MIDDLE SCHOOL

	Academic Index			
	1992-93	1993-94	1994-95	1995-96
Reading	41.5380	39.9980	45.3300	38.8920
Mathematics	22.7676	30.6680	34.5386	45.5574
Science	18.4600	8.0000	24.0000	15.5560
Social Studies	36.9160	36.0020	30.6680	43.3360
Arts & Humanities	20.0000	12.0000	20.6620	21.6620
Prac. Liv./Voc. Studies	6.6680	14.2840	24.6620	3.3320
Writing Portfolio	6.1520	41.4300	49.3300	37.7760

Total Academic Index	23.2	28.2	34.4	32.3
Number Tested	13	15	15	18
Accountability Index	34.8	38.8	44.0	42.2

Noncognitive Indicators				
	1992-93	1993-94	1994-95	1995-96
Attendance Rate	94.54	94.26	94.26	93.05
Dropout Rate	0.00	2.33	2.33	0.00
Retention Rate	8.33	7.14	7.14	6.67
Successful Transition to Adult Life				
Noncognitive Index	94.4840	94.3820	94.3820	94.5520

Year Data Are Collected					
1992-93	1993-94	1994-95	1995-96		
▨	▨			Accountability Cycle II Baseline Index *	36.9
▨	▨			Accountability Cycle II Improvement Goal *	43.2
		▨		Accountability Cycle II Growth Index 1994-95	44.0
			▨	Accountability Cycle II Growth Index 1995-96	42.2
		▨	▨	Accountability Cycle II Combined Growth Index *	43.0

* Weighted Average for 2 years.

Because this school has more than one accountability grade, the accountability index depends on the indices of both grade 8 and grades 11/12. Your school's combined accountability index is being sent from the Department of Education under separate cover.

RUN DATE: 10/11/96

TM030643

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

EFF-089 (9/97)